GESTION / FINANCE

Les Essentiels Métiers

Appliquer la data science à l'assurance

David Dubois • Aurélie Féderlé
Voahirana Ranaivozanany

L'ARGUS de l'assurance ÉDITIONS

Sommaire

Introduction ... 9

1. Les clés de lecture du monde de la donnée

Comprendre l'émergence des concepts big data et data science ... 15
Réglementer la collecte, le traitement et l'utilisation des données pour asseoir la confiance ... 22
Intégrer l'éthique comme composante fondamentale du traitement des données ... 28
Rendre un algorithme interprétable, la clé pour une prise de décision éclairée ... 35
Exploiter la donnée pour en exprimer la valeur ... 37

2. Comment constituer sa base de données ?

Bien cerner la base de données souhaitée ... 43
Cartographier les bases de données disponibles en interne ... 48
Comprendre les remontées d'informations ... 52
Recenser les informations internes non structurées ... 55
Enrichir par la recherche de données externes ... 56

3. Exploiter les mégadonnées à l'aide de l'apprentissage automatique

Nettoyer les données : le préalable à toute analyse ... 59
Comprendre et identifier les liens entre variables : le feature engineering ... 67
Sélectionner les variables pertinentes par le processus dit feature selection ... 70
Apprendre à l'aide des données : l'apprentissage machine ... 75
L'apprentissage automatique profond (Deep Learning) ... 84

4. Applications des domaines de la data science à l'assurance : quelques cas d'usages

Exemple d'application n° 1 : Amélioration de la tarification santé par l'intégration du comportement de consommation ... 91
Exemple d'application n° 2 : comportement d'arbitrage en assurance vie ... 94
Exemple d'application n° 3 : de la data science en assurance téléphonie mobile ... 97
Exemple d'application n° 4 : assurance de flotte automobile ... 100

5. La transformation des métiers induite par le développement de la data science

Comprendre les enjeux de la digitalisation des organismes d'assurance sur les métiers ... 105
La révolution numérique au service de la fonction actuarielle ... 110
L'émergence des data scientists ... 111
Vers un actuariat augmenté ... 113

Bibliographie ... 119

Index alphabétique ... 123

Comment constituer sa base de données ?

En raison de l'inversion du cycle de production en assurance (les coûts de production ne sont connus qu'une fois la sinistralité évaluée), les organismes d'assurance ont toujours disposé de données en quantité sur les assurés et sur les risques. Ils ont constitué des bases de données avec une large profondeur historique sur des risques souscrits depuis longtemps. L'objectif de ce chapitre sera de proposer une démarche pour constituer une base de données qui permettent d'appliquer les outils de la data science.

Bien cerner quelle est la base de données souhaitée est un préalable indispensable. Le sujet sera ensuite d'identifier parmi les données disponibles celles qui vont permettre de créer de la valeur pour l'entreprise, puis de les réorganiser en vue de les exploiter conjointement en appliquant des approches de data science. Le sujet sera aussi d'enrichir ces données de portefeuilles par des données qui jusqu'alors n'étaient pas utilisées, ces données supplémentaires pouvant être internes ou externes à l'entreprise.

1. Bien cerner la base de données souhaitée

1.1 Les assureurs sont de grands consommateurs historiques de données

Les données ont toujours été au centre des activités de l'organisme d'assurance quels que soient la branche d'assurance ou les produits

considérés. Dans un contrat d'assurance, l'assureur prend l'engagement d'indemniser l'assuré dans le cas où le risque couvert se réalise. Il s'agit, par exemple, de verser un capital en cas de décès, de prendre en charge des frais de soin de santé en cas de maladie, de verser des indemnités journalières ou rentes d'invalidité en cas d'arrêt de travail, d'assurer la réparation de l'habitation en cas d'incendie ou encore de réparer le véhicule en cas d'accident.

Donner un prix à un contrat d'assurance revient à mesurer le coût probable pour l'organisme d'assurance en considérant conjointement les montants des garanties souscrites et la probabilité de survenance du risque. Pour ce faire, l'organisme d'assurance s'appuie sur un ensemble d'informations concernant la communauté d'assurés ciblée par le contrat mais aussi des informations individuelles pour affiner sa tarification pour un assuré considéré.

EX

Ainsi, en assurances de personnes, des informations telles l'âge, l'état civil, la catégorie socio-professionnelle ou encore les données sur l'état de santé sont alors utilisées. Dans le cas des assurances aux biens, l'organisme assureur demande également des informations sur les biens couverts telles la marque et la date de première immatriculation pour les véhicules ou encore la localisation exacte des habitations.

Dès la mise en place d'une couverture, l'assureur recueille donc une quantité importante d'informations, indispensables à la contractualisation, ces informations servant à affiner la tarification mais aussi à garantir l'état du risque au moment de la souscription. Durant la vie des contrats d'assurance, l'organisme d'assurance aura connaissance de l'évolution du risque assuré ce qui lui permettra de procéder à une évaluation régulière de ses engagements au cours du temps et de constituer les provisions conformément aux exigences réglementaires. Les assurés sont donc tenus d'informer l'organisme d'assurance de toutes les modifications pouvant impacter les risques couverts. Ces informations, complétées des informations relatives aux modifications de garanties ou demandes de prestations, viennent enrichir au fur et à mesure la masse de données déjà recueillies dès la souscription.

Concernant les produits d'assurance vie (épargne), pour les mêmes motivations et obligations réglementaires, les assureurs collectent également des données sur les assurés et leurs contrats. Ces données personnelles,

les montants et dates des versements, arbitrages, rachats et avances ou encore supports d'investissements sont la matière sur laquelle l'organisme assureur effectue ses travaux. La réglementation sur la lutte contre le blanchiment de capitaux et le financement du terrorisme (LCB-FT) se traduit pour les assureurs par des exigences encore plus fortes sur la connaissance de leurs clients, ainsi que sur les origines ou destinations des fonds lors des opérations réalisées tout au long de la vie des contrats. Les clients sont connus de manière de plus en plus détaillée et les informations recueillies sont de plus en plus fiables. Des procédures adéquates ont été mises en place par les organismes assureurs accroissant encore les données qu'ils collectent.

Exemple de procédure LCB-FT

Source : « Le machine learning en finance – vers une véritable révolution », A. Caiazzo, C. Na Champassak, G, Floch, H. Sale, Juillet 2020

1.2. Pourquoi parler d'une base de données à constituer ?

Avec autant d'informations déjà disponibles, on pourrait s'interroger sur le besoin de constituer une nouvelle base de données.

▶ Ce que les organismes d'assurance font traditionnellement avec les données collectées

Schématiquement, les données sont utilisées pour déterminer les bases tarifaires, établir des tables d'expérience, réaliser les travaux d'inventaires et les comptes de résultats par portefeuille, suivre l'équilibre technique et la rentabilité et évaluer les besoins en capital réglementaire. Il s'agit principalement de travaux actuariels, réglementaires et comptables permettant aux organismes d'assurance de satisfaire à leurs obligations. Les données sont également utilisées pour les travaux actuariels.

Dans la majorité des cas, des requêtes et des processus de qualification automatisés des données ont été mis en place pour ces activités. Pour chaque inventaire, les services actuariels vont par exemple recevoir directement la liste des assurés en portefeuille pour les travaux d'estimation

des provisions tête par tête (par exemple les calculs des provisions incapacité / invalidité à constituer). Le fichier comporte les variables nécessaires aux calculs qui seront par la suite intégrées dans un outil de calcul de provisions : date de naissance, date de début de l'arrêt, date de fin de garantie, montant de prestation à servir…

Autre exemple, pour la mise en place des tables d'expérience qui doivent être révisées régulièrement, les données vont être obtenues de la part des services de gestion ou en faisant tourner une requête à disposition. Les expressions de besoins ont été effectuées antérieurement et les données sont extraites de manière analogue chaque fois que les tables sont à établir (variables nécessaires et données sur les années d'observation choisies). De telles approches industrialisées ont été adoptées pour optimiser les temps de traitement et pour limiter les risques opérationnels liés à un oubli d'une partie des données par exemple, l'expertise du service actuariel vient en regard complémentaire de validation et pour l'analyse.

Au sein des directions techniques des organismes d'assurance, les tâches sont généralement réparties sur des collaborateurs autonomes sur leurs dossiers. Les organisations ont instauré des silos sur les différents périmètres en créant par exemple des équipes par famille de produits (épargne, prévoyance, santé, auto, MRH…) ou par type d'activités (validation du chiffre d'affaires et des prestations, calculs des provisions arrêt de travail, calculs des provisions santé…) ou autres. Ces découpages ont été choisis par l'entreprise selon les axes d'analyse qui répondent au mieux à ses préoccupations propres. Avant les travaux de consolidation nécessaires, les calculs et analyses sont donc souvent effectués de manière dissociée par des équipes très spécialisées.

Ces approches répondent très bien aux besoins traditionnels des assureurs. Elles incluent néanmoins des limites pour la data science. En effet, les équipes maîtrisent les données nécessaires à leurs propres travaux mais ne connaissent pas toujours l'exhaustivité des données existant sur leurs périmètres. D'ailleurs, la connaissance, par chacun, de l'intégralité des données sur l'entreprise est encore limitée plus fortement.

▶ Quelle base de données pour la data science ?

L'intérêt de la data science réside dans sa capacité à traiter des volumes importants d'informations. Elle permet de traiter un grand nombre d'occurrences mais également un grand nombre de variables. En d'autres termes, les portefeuilles peuvent être analysés à des mailles très fines, avec une grande profondeur d'historique, et avec des caractéristiques très détaillées balayant plusieurs volets.

Les avancées technologiques de ces dernières années ont rendu possibles une réelle exploitation de la data science. Les capacités de stockage de données sont de plus en plus importantes et les puissances et temps de calculs ont beaucoup évolué. Les données sont devenues accessibles et disponibles plus rapidement et plus facilement. Les outils de la data science peuvent alors être plus largement diffusés et utilisés. En effet, si les méthodes scientifiques sous-jacentes existent depuis longtemps, elles n'ont pas pu servir pleinement avant ces évolutions informatiques notamment.

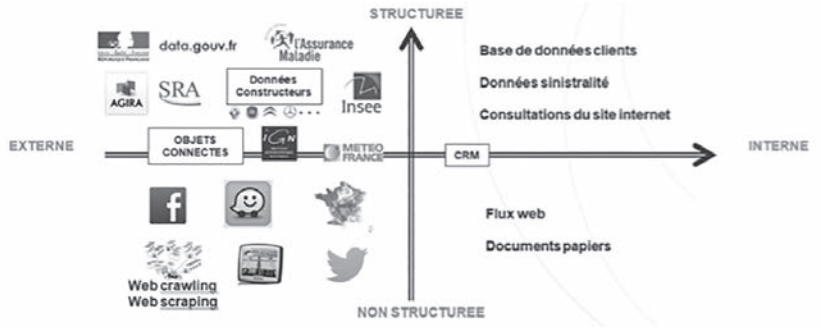
Pour que la data science prenne tout son sens, il conviendrait toutefois de disposer d'une structure et d'une organisation des données permettant de profiter de ses atouts. Les données disponibles sur chaque périmètre doivent, notamment, être connues. Le stockage doit faciliter une vision globale sur l'ensemble des données de l'entreprise, tous périmètres rassemblés, en cassant les silos existants habituellement. Le stockage doit également permettre leurs exploitations conjointes.

Par ailleurs, les données non structurées sont toutes aussi importantes que les données déjà organisées en base de données dans les entreprises. Ces données éparpillées aujourd'hui dans les différents services de l'entreprise doivent être organisées pour optimiser les apports de la data science, voire pour fiabiliser les résultats obtenus.

La data science rend possible une utilisation en masse des données externes, qu'il s'agisse de données structurées ou non. Une mise à jour régulière des données est également possible et permettra d'être réactif et proactif en tenant continuellement compte des évolutions de contexte ou de tendance.

De manière globale, les données utilisées par la data science ont deux axes de différenciations majeurs : données internes ou externes, et données structurées ou non structurées comme illustrées ci-dessous. La base de données à constituer devrait permettre d'inclure l'ensemble de ces données et de les travailler ensemble.

Sources de données exploitables en assurance



Source : « Big data, machine learning et objets connectés : les nouveaux challenges de l'assurance automobile », Forsides, 2015

2. Cartographier les bases de données disponibles en interne

2.1 Recenser toutes les bases de données pour ouvrir le champ des possibles

Les données disponibles sur les portefeuilles assurés sont un réel patrimoine de l'organisme assureur. Elles incluent les caractéristiques des assurés et retracent leurs comportements en tant qu'assuré, leurs choix et leurs décisions concomitamment à l'évolution de leurs contrats et de leurs risques. L'ambition est d'extraire les richesses au sein de ces données sur les assurés, pour les faire « parler », et en enrichir les approches de pilotage et de création de valeur de l'organisme d'assurance. Il s'agit de richesses non exploitées par les méthodes statistiques standards.

Aujourd'hui, les projets data science se situent principalement autour de trois axes : le développement commercial, le pilotage technique et la gestion avec en filigrane le respect de la conformité des opérations. Toutefois, ils peuvent être de natures très variables comme illustré par le tableau ci-dessous.

Nature des projets data science en assurance

	DOMAINES	TYPES DE PROJETS
AXE DÉVELOPPEMENT	OPTIMISATION DE LA VALEUR CLIENT	• Analyse de l'élasticité aux majorations • Analyse du risque de résiliation • Anticipation des évolutions de comportement
	CRM, MARKETING ET CIBLAGE	• Meilleure connaissance des clients et optimisation globale du parcours client • Ciblage des campagnes marketing en vue d'en maximiser l'efficacité • Optimisation du taux de concrétisation des devis
	OFFRE ET SERVICES	• Conception de nouveaux services rendus possibles par les nouvelles données • Intégration des nouvelles technologies dans les offres actuelles • Interaction en temps réel
AXE TECHNIQUE	SUIVI TECHNIQUE	• Identification des segments déficitaires et modulation des majorations tarifaires • Anticipation des évolutions du risque • Aide à la détermination du coût ultime de la sinistralité
	TARIFICATION	• Analyse de l'élasticité au prix d'achat et intégration dans le « competitive pricing » • Modification du tarif par l'intégration de « scorings » multicritères • Modification du tarif par l'intégration de données externes nouvelles
AXE GESTION	FRAUDE	• Identification des comportements atypiques • Aide à la décision sur cas litigieux et aide à la souscription • Identification des informations déclaratives falsifiées
	SINISTRES	• Estimation du degré d'urgence et priorisation des dossiers à traiter • Contrôle de la bonne gestion et de la correcte liquidation des dossiers • Aide à la détection de risques d'erreurs dans le paiement des dossiers

Source : Forsides

Les données utilisées peuvent être très différentes d'un projet à l'autre avec (liste non exhaustive) :
- des données de gestion ou des données comptables ;
- des données sur les assurés ou sur les sinistres ;
- des données du moment ou intégrant l'historique ;
- des données provenant des systèmes d'informations de l'assureur ou transmises par les partenaires et les délégataires ;
- des données issues des systèmes d'information ou des tableurs excel ;
- ou encore des données mises à jour périodiquement selon une fréquence décidée.

Pour élargir au maximum l'éventail des applications potentielles, un recensement de l'ensemble des bases de données existantes doit être réalisé. La recherche est à mener auprès de tous les départements de l'entreprise. Un travail de vulgarisation sur les apports et les prérequis de la data science pourrait être nécessaire pour faire adhérer les collaborateurs, surtout lorsque des silos organisationnels sont trop bien établis.

Ce premier travail de recensement est essentiel pour la mise en place d'une base de données avant tout travail. Pour cette étape, il convient de

ne s'arrêter sur aucun a priori concernant l'utilité, la pertinence ou la qualité des données. Aucun tri ni choix a priori ne doit être réalisé. Le principe est de commencer par tout répertorier de manière la plus exhaustive, de manière à ouvrir au maximum le champ des possibles.

2.2 Identifier les sources de données et dénombrer leurs caractéristiques

Le recensement des bases de données est une étape primordiale qui permettra de bien identifier les données à disposition pour pouvoir en profiter pleinement et efficacement. Pour ce faire, la connaissance des schémas de collecte et des systèmes d'information sont les prérequis essentiels.

▶ Connaître les variables et les documenter

La connaissance extensive des données reste encore trop limitée dans les organismes d'assurance. Chaque service voire chaque utilisateur pris isolément connaît et maîtrise les données nécessaires à ses travaux sur son périmètre, sans pour autant connaître l'exhaustivité de celles-ci sur l'ensemble des périmètres de l'entreprise. La connaissance des périmètres autres que le sien est limitée encore plus fortement, peu de personnes ayant une visibilité transversale sur l'ensemble des données de l'entreprise. Il est même facile de dire que dans certaines structures, personne n'a cette visibilité.

Faire abstraction des variables non connues réduit d'autant la valeur patrimoniale des « données ». Pour capitaliser au mieux sur une base de données, il est important d'identifier à quoi correspond chacune de ses variables. Une partie des variables est utilisée couramment et donc bien connue par les équipes qui en font appel. Certaines pourraient être reconnues spontanément même si non utilisées au quotidien. Pour le reste, si nécessaire, les spécifications écrites lors de la mise en place des bases de données, complétées par des échanges avec les différents métiers de l'entreprise pourraient aider.

En tout état de cause, une documentation centralisée des variables permettrait d'y accéder plus facilement et pérennisera cette connaissance. Ce descriptif pourrait partir d'un existant, complété ou ajusté selon les besoins.

Solvabilité 2, avec ses exigences sur la qualité des données, a conduit à initier ces travaux au sein des organismes d'assurance. Des premiers dictionnaires ont été établis, sur les données nécessaires aux calculs de provisions *best-estimate* et de besoins en capital.

Appliquer la data science à l'assurance

David Dubois

Actuaire certifié IA, expert ERM et qualifié CERA, titulaire d'un certificat Business sustainability management Cambridge, d'une certification Design sprint, d'un certificat Marketing digital ESCP Europe, d'un Executive MBA de l'ESCP Business School et d'un DEA Economie mathématique et économétrie, David Dubois est directeur des partenariats et synergies du groupe Prevoir, enseignant et formateur. Il a été président de l'Institut des actuaires de 2016 à 2020.

Aurélie Féderlé

Diplômée de l'Institut de statistique de l'université de Paris et titulaire du master Ingénierie statistique et data science de Sorbonne université, elle est data scientist chez Dataltist (groupe Forsides).

Voahirana Ranaivozanany

Actuaire certifiée IA, expert ERM et qualifiée CERA, elle est Fondée de pouvoir chez Forsides, cabinet de conseil en actuariat, après plus de 20 ans en compagnie. Elle est chargée d'enseignement en actuariat depuis 2013 et membre du jury de la formation Expert ERM - CERA.

Nous vivons une période de forte numérisation de notre monde où tout est susceptible de devenir une donnée numérique. Ces données pléthoriques et ces nouveaux outils offrent aux organismes d'assurance la possibilité d'être plus performants et plus réactifs pour, *in fine,* mieux servir leurs clients. C'est tout l'objectif de la data science qui vise à extraire la substantielle connaissance des données pour orienter les décisions et les prévisions.

Pour les organismes d'assurance, la data science porte les gênes d'une transformation dans leurs modèles d'affaires depuis la conception de l'offre au règlement du sinistre en passant par la gestion de la relation client. Automatisation des processus, amélioration de la connaissance client, réduction du taux d'attrition ou encore repérage des comportements de fraude ne sont que quelques sujets sur lesquels les techniques de la data science peuvent apporter des améliorations sensibles voire radicales au service de la création de valeur.

Ponctué d'exemples pratiques, cet ouvrage se veut une introduction à l'univers complexe des big data et de la data science appliqué au secteur de l'assurance.

ISBN 978 2 35474 362 8

Photo de couverture : Fotolia

Les Essentiels Métiers