Qualité de données et techniques de détection d'outliers

Alexandre Boumezoued, Directeur R&D Milliman & Marine Habart, AXA Life and Health Chief Actuary

Webinaire Milliman

04 FEVRIER 2021

Milliman

Qualité des données et détection des outliers

Introduction

Le rôle d'un actuaire est de quantifier et de résoudre les potentiels dysfonctionnements le long de la chaîne de production. Les déficiences peuvent être divisées en quatre champs :

- Qualité des données : incertitude sur les variables d'entrée du modèle. L'erreur contenue dans chaque variable peut provenir d'une saisie incorrecte de données ou résulter d'erreurs provenant d'un autre modèle. Cette incertitude réside également dans le respect ou non des hypothèses du modèle choisi.
- Pertinence du modèle : incertitude contenue dans le modèle. Le modèle peut être mal choisi vis-à-vis de la situation réelle étudiée, mal implémenté ou peut manquer de fondements théoriques.
- Instabilité du modèle : il s'agit de déterminer si le modèle n'est pas trop sensible à certaines hypothèses prédéfinies.
- Interprétation du modèle : dans ce dernier champ, on considère l'erreur causée par une mauvaise interprétation ou une mauvaise utilisation des résultats du modèle. Cela repose sur une bonne compréhension de l'interaction et de l'impact de toutes les variables du modèle sur le calcul final.

Qualité des données et détection des outliers

Zoom sur la qualité des données

- Il est fréquent que certaines observations s'écartent significativement du reste de la base de données.
- Ces points de données sont appelés « outliers » ou valeurs aberrantes. Il peut s'agir :
  - De **points erronés** qui doivent être traités (ex : mauvaise qualité des données, mauvais encodage)
  - D'**observations exceptionnelles** (ex : réalisations extrêmes dans un échantillon, évènements catastrophiques)
  - D'observations pertinentes (non erronées) qui se distinguent des autres points et qui peuvent contenir **une information fiable, importante et inattendue** (ex : comportement anormal ou frauduleux, écart de risque)
- Une fois la nature des outliers identifiée**, il faut décider s'ils doivent être corrigés ou retirés** du jeu de données :
  - Rectifier la valeur de ces observations
  - Les retirer du jeu de données (potentiellement les remplacer)
  - Les conserver s'ils constituent des points de données pertinents

Isolation forest

Aperçu de la méthode

- **Objectif** : déterminer un classement qui reflète le caractère « outlier » de chaque point de données
  - Trier les observations selon leurs longueurs de chemin ou leurs scores d'anomalie
  - Les outliers sont les observations avec les scores d'anomalie les plus élevés
- **Isolation Tree (iTree)** : arbre binaire où chaque nœud de l'arbre possède exactement zéro ou deux fils.
- **Algorithme d'Isolation Forest (iForest)** : algorithme d'apprentissage non supervisé inspiré des forêts aléatoires
  - Non supervisé : les données ne sont pas étiquetées
  - Pas nécessaire de définir le profil d'une donnée normale ni de calculer des distances entre points
  - Construit un ensemble d'arbres aléatoires grâce à un mécanisme appelé « isolation », un processus de partitionnements itératifs et aléatoires pour séparer les outliers des points normaux
  - Se base sur l'observation **que les outliers ont une probabilité plus élevée d'être isolés des autres points en moins d'étapes** que les points normaux.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. In 2008 Eighth IEEE International Conference on Data Mining, pages 413-422. IEEE.

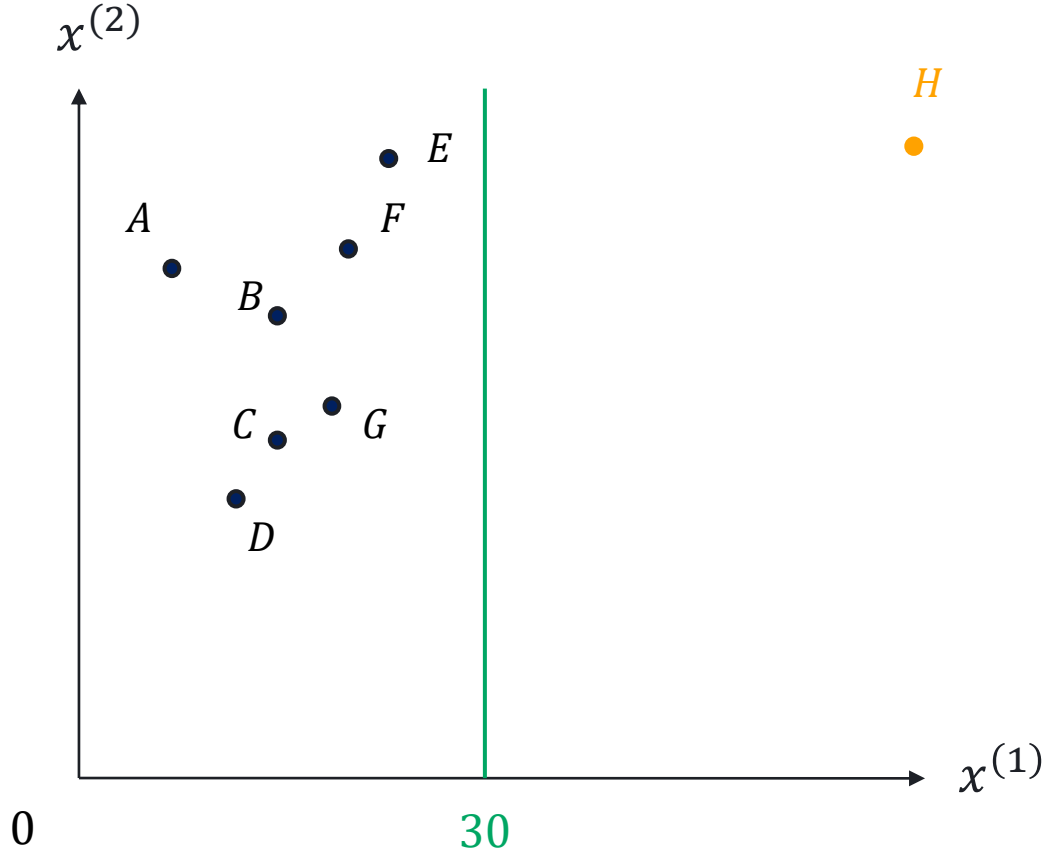
Hyunsu Kim and Michael Leitschkis. (2019). See the forest for the trees. Milliman white paper available at https://www.milliman.com/en-gb/insight/see-the-forest-for-the-trees

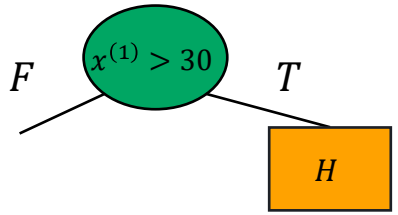
R package solitude: An Implementation of Isolation Forest. Komala Sheshachala Srikanth. https://cran.r-project.org/package=solitude

Isolation forest

Exemple d'Isolation Tree en dimension 2

- Le point H (outlier) est isolé en seulement une étape
- Plus d'étapes sont nécessaires pour isoler les autres points

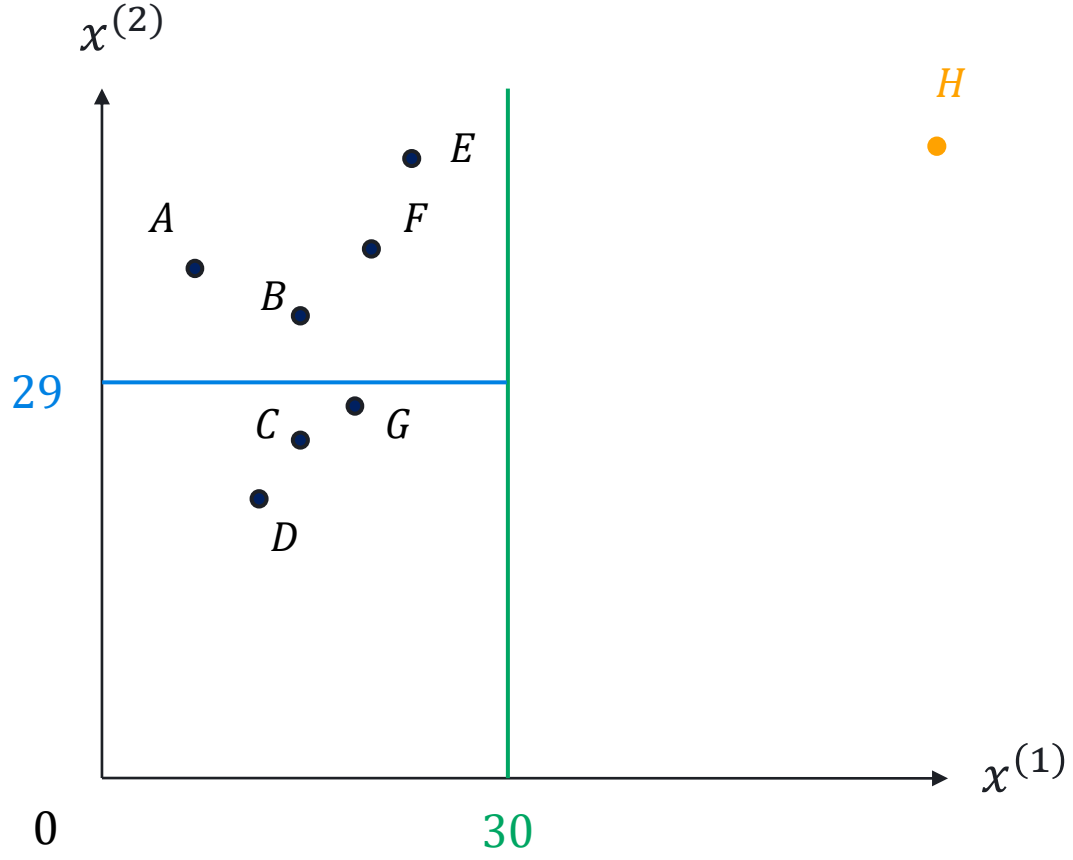


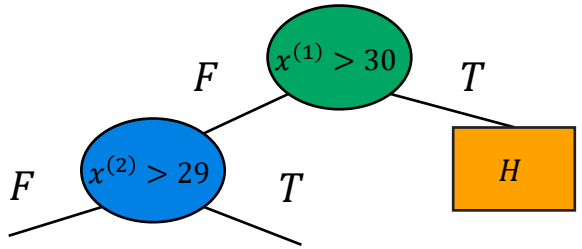


Isolation forest

Exemple d'Isolation Tree en dimension 2

- Le point H (outlier) est isolé en seulement une étape
- Plus d'étapes sont nécessaires pour isoler les autres points

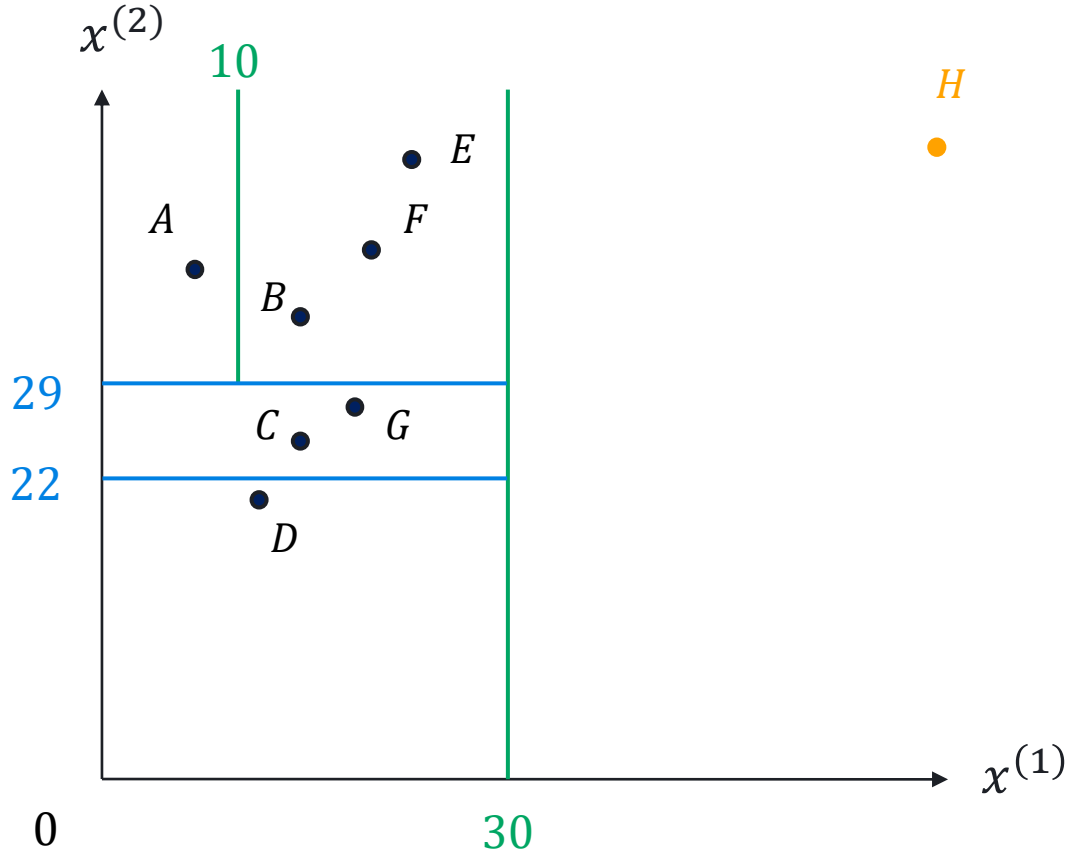


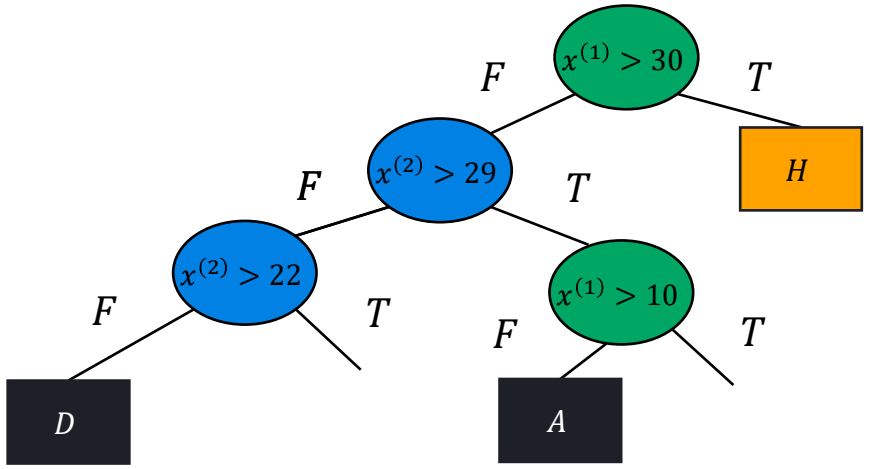


Isolation forest

Exemple d'Isolation Tree en dimension 2

- Le point H (outlier) est isolé en seulement une étape
- Plus d'étapes sont nécessaires pour isoler les autres points

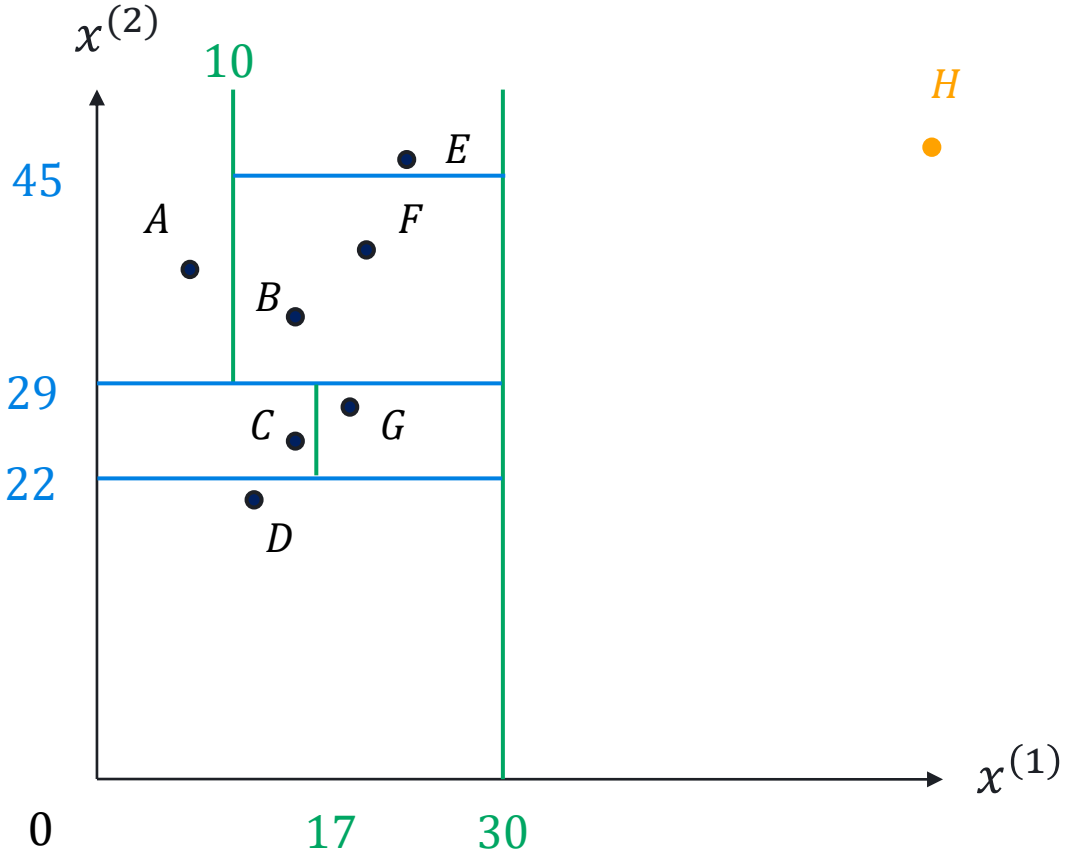


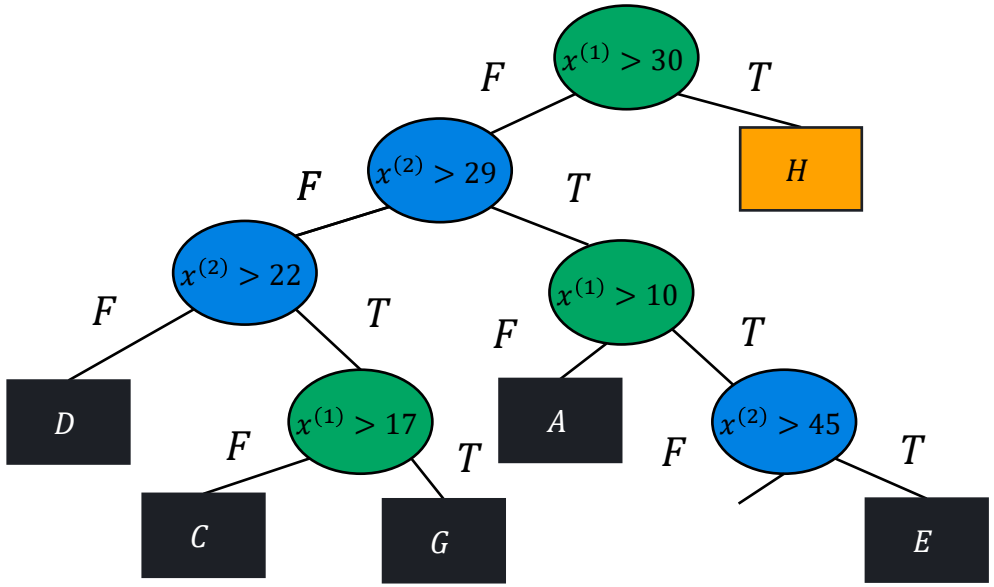


Isolation forest

Exemple d'Isolation Tree en dimension 2

- Le point H (outlier) est isolé en seulement une étape
- Plus d'étapes sont nécessaires pour isoler les autres points

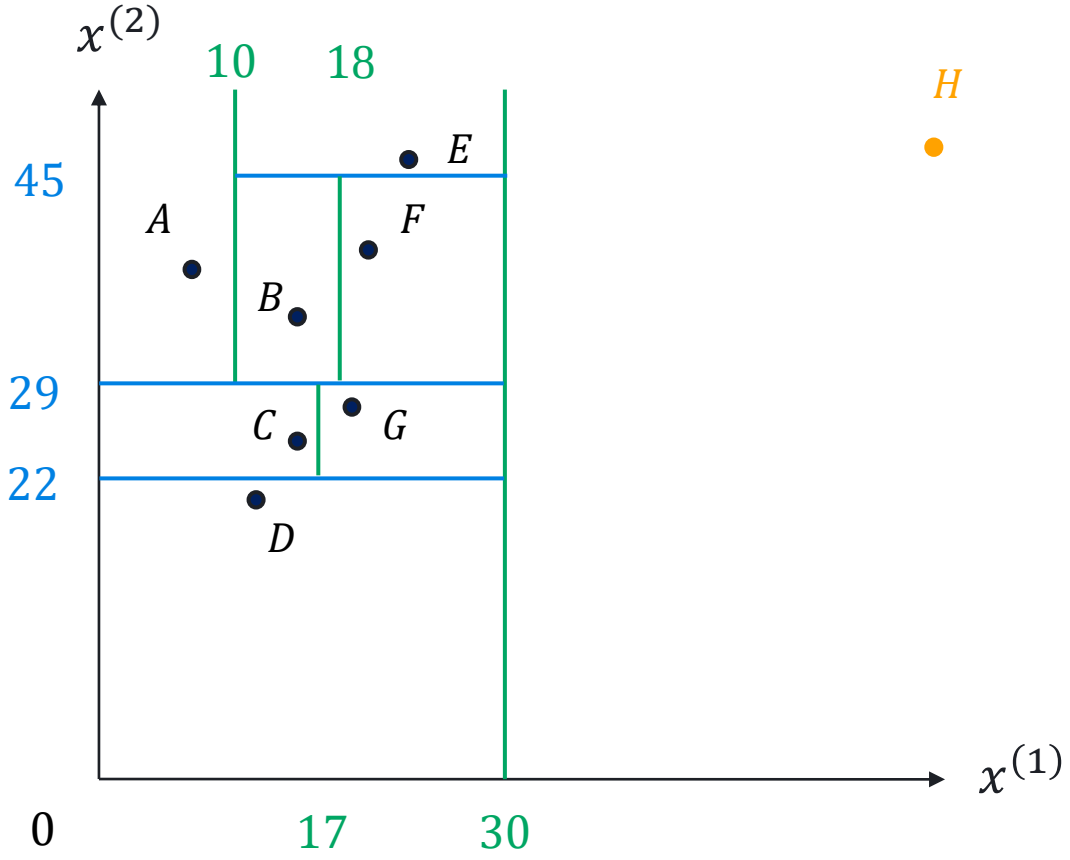


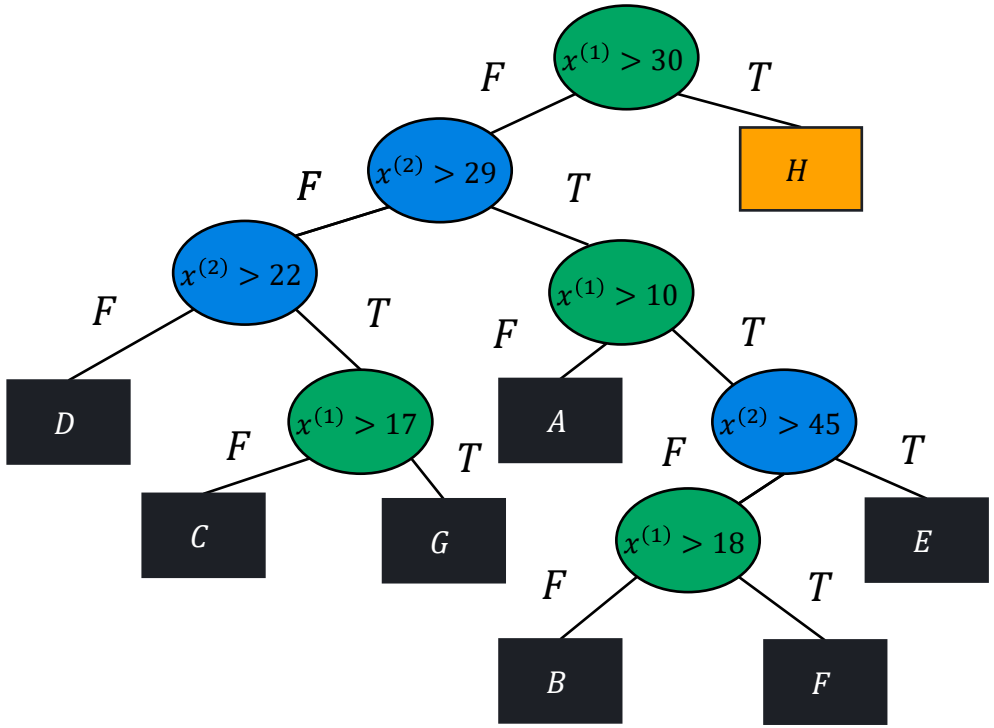


Isolation forest

Exemple d'Isolation Tree en dimension 2

- Le point H (outlier) est isolé en seulement une étape
- Plus d'étapes sont nécessaires pour isoler les autres points

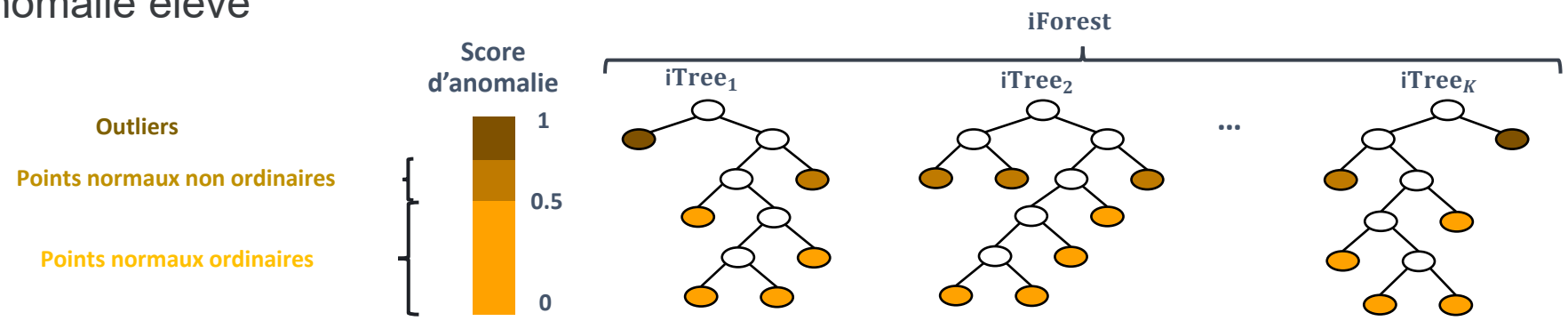




Isolation forest

De l'iTree à l'iForest

- **Apprentissage ensembliste : génération de plusieurs iTrees → iForest**
  - La longueur de chemin d'une observation est calculée comme la somme :
    - du nombre total de séparations nécessaires pour l'isoler
    - d'un terme d'ajustement, ajouté lorsque l'observation se situe à un nœud externe. (Prend en compte un sous-arbre hypothétique au-delà d'une longueur limite ℓ → économie de calcul)
  - Détermination de la longueur moyenne de chemin $\overline{h(X_i)}$ pour chaque observation X_i
- **Calcul des scores d'anomalie**
  - Normalisation par la longueur moyenne de chemin (universelle) L dans un arbre binaire
  - Score d'anomalie : $s(X_i) = 2^{-\frac{\overline{h(X_i)}}{L}} \in [0,1] \Rightarrow$ faible longueur moyenne de chemin = score d'anomalie élevé



Isolation forest

Swamping et Masking

- Le *swamping* et le *masking* sont deux problèmes courants rencontrés en détection d'anomalies
  - ***Swamping* :** des points de données normaux sont détectés à tort comme des outliers dans le cas où un grand nombre de variables ne sont pas informatives sur la nature d'outlier (la séparation via ces variables n'est alors pas pertinente)
  - ***Masking* :** lorsque trop d'outliers sont présents dans le jeu de données, les règles de séparation ne sont plus efficaces car trop d'itérations sont nécessaires.
- Ces deux problèmes proviennent notamment d'un nombre excessif d'observations
- Solution de l'algorithme iForest : **le sous-échantillonnage**
  - Limite la taille des données pour mieux isoler les exemples d'outliers
  - Chaque iTree peut être spécialisé, chaque sous-échantillon contenant un ensemble différent d'outliers
- On montre que la détection d'outliers via l'algorithme iForest fonctionne mieux lorsque l'on a recours au sous-échantillonnage

Isolation forest

Les caractéristiques de l'algorithme

- **Avantages** :
  - Méthode non supervisée : ne nécessite pas l'identification d'outliers par avis d'experts
  - Ne nécessite pas de modèle (l'objectif n'est pas de modéliser des instances normales)
  - Fournit une hiérarchie en attribuant un score d'anomalie pour chaque observation
  - Utilise des échantillons de tailles relativement petites obtenus à partir de grandes bases de données pour déterminer une fonction de détection d'outliers
  - L'algorithme peut être entraîné une fois puis réutilisé sans coût de calcul
  - La complexité en temps est linéaire et l'algorithme requiert un faible besoin de mémoire
    - En utilisant le sous-échantillonnage
    - En évitant de construire des arbres dépassant la longueur limite ℓ
  - Résout les effets de *swamping*/*masking* grâce au sous-échantillonnage
- **Inconvénients :**
  - Nécessite un travail sur les données pour ajuster le format des variables
  - Des paramètres doivent être fixés : taille de sous-échantillonnage, longueur maximale d'un arbre, nombre d'arbres…
  - La valeur limite de score pour qualifier les « outliers » reste un jugement d'expert

Applications de l’isolation forest

Base de données

Données : ≈3500 observations, 2 types de variables

9 variables continues

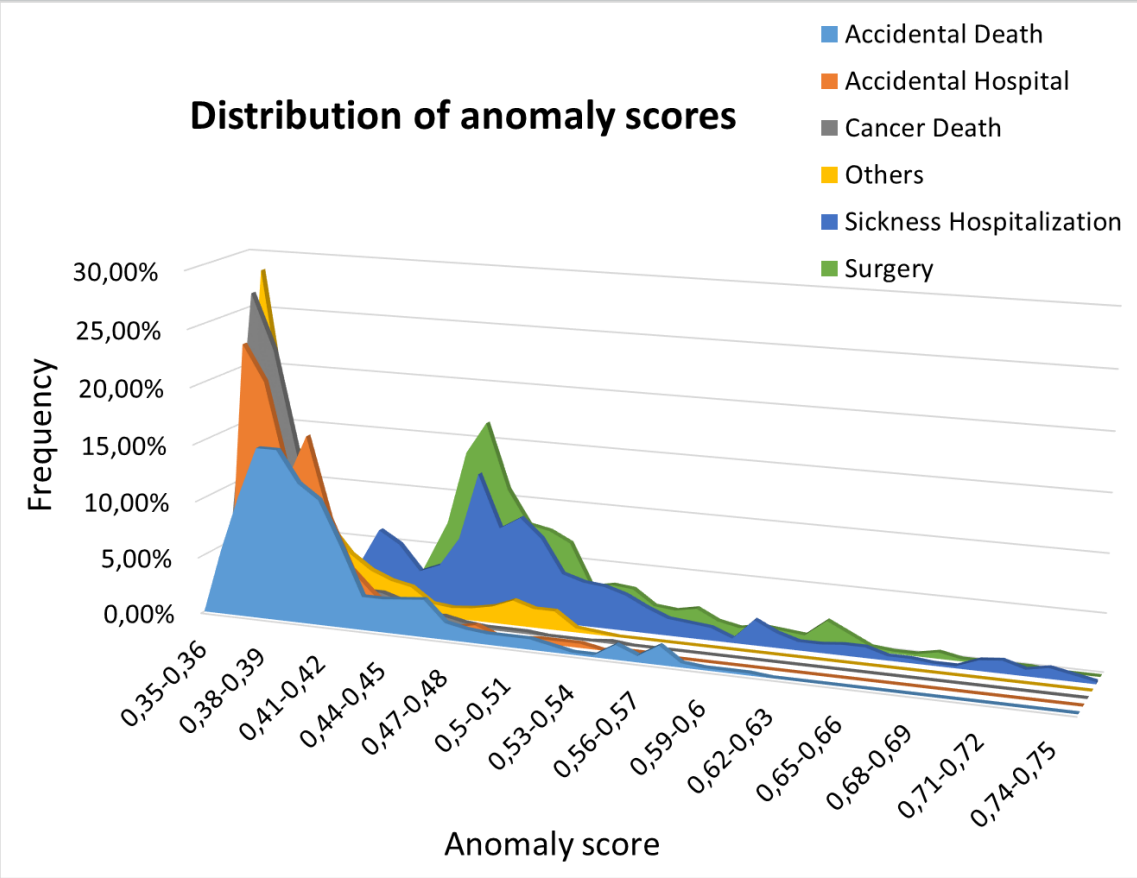
- Année d’observation t
- In-Force total
- Nombre de sinistres déclarés
- Total de sinistres attendus
- Total de sinistres effectifs
- Best Estimate
- Facteur d’ajustement
- Exposition totale
- Ratio $\frac{\text{Total de sinistres effectifs}}{\text{Exposition totale}}$

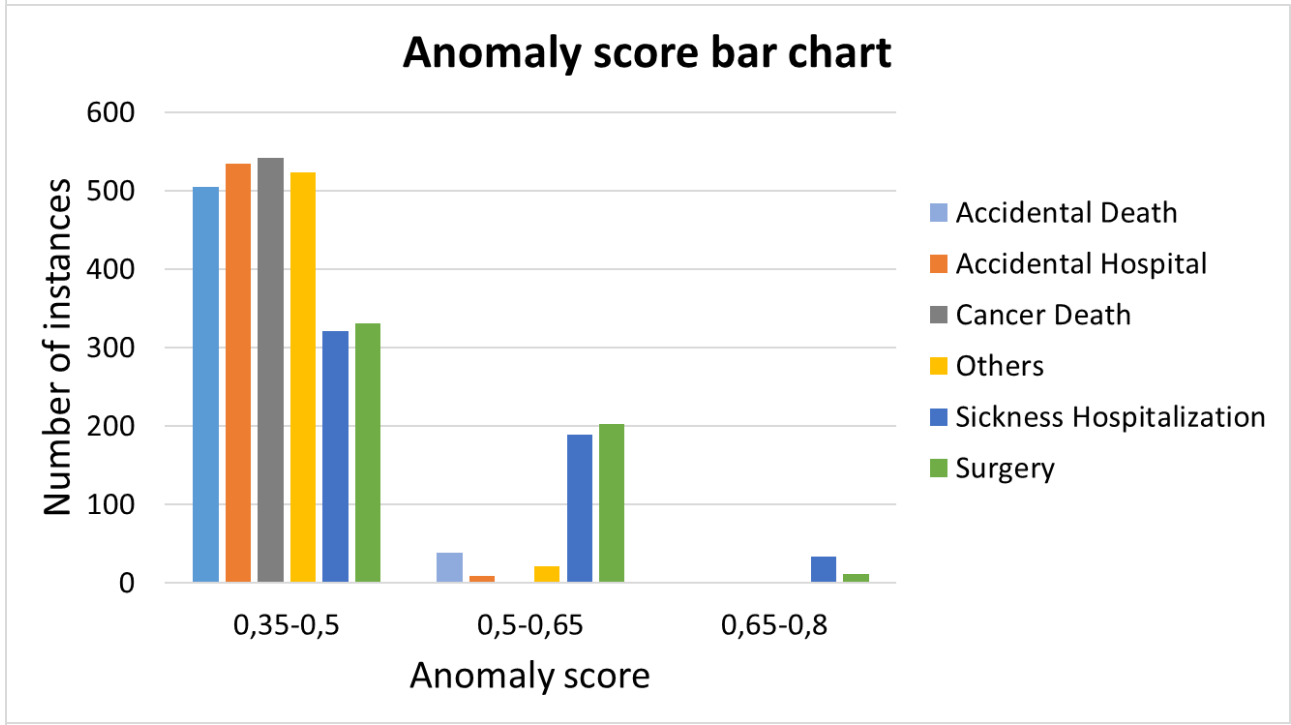
3 variables catégorielles

- Segment : Accidental death, Accidental hospitalization, Cancer death, Sickness hospitalization, Surgery, Other
- Homme/Femme
- Ancienneté

Applications de l'isolation forest

Distribution des scores d'anomalies par segment





Applications de l’isolation forest

Top 25 des scores d’anomalie par ordre décroissant

1er type d’outliers :

Segment « Sickness Hospitalization » avec une ancienneté commune (la plus élevée)

2eme type d’outliers :

Segment « Sickness Hospitalization » avec une année de souscription donnée (2005 et 2006)

3eme type d’outliers :

Segment « Surgery »

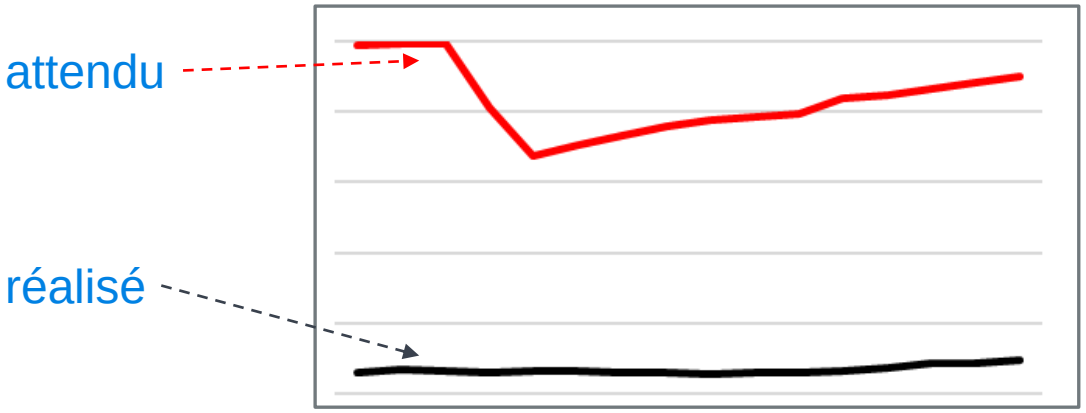
Segmentation Group	iForest score
Sickness Hospitalization	0,752
Sickness Hospitalization	0,748
Sickness Hospitalization	0,744
Sickness Hospitalization	0,740
Sickness Hospitalization	0,740
Sickness Hospitalization	0,738
Sickness Hospitalization	0,736
Sickness Hospitalization	0,735
Sickness Hospitalization	0,731
Sickness Hospitalization	0,728
Surgery	0,726
Sickness Hospitalization	0,724
Sickness Hospitalization	0,722
Sickness Hospitalization	0,717
Sickness Hospitalization	0,717
Sickness Hospitalization	0,717
Sickness Hospitalization	0,714
Sickness Hospitalization	0,713
Surgery	0,712
Sickness Hospitalization	0,712
Sickness Hospitalization	0,709
Sickness Hospitalization	0,708
Sickness Hospitalization	0,707
Sickness Hospitalization	0,706
Surgery	0,706

Applications de l’isolation forest

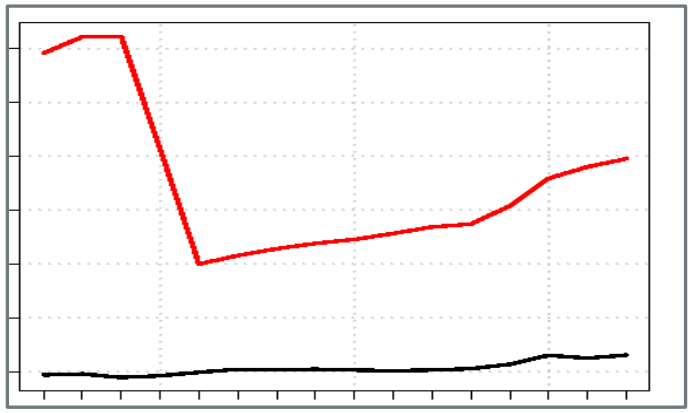
Outliers de Type 1 – rupture de sinistres attendus

- Segment « Sickness Hospitalization» : Le montant agrégé des sinistres attendus chute brutalement sur une période d’observation donnée, tandis que le montant agrégé des sinistres réalisés reste stable sur cette période
- Les anciennetés les plus élevées sont détectées comme des outliers par l’algorithme
- L’analyse graphique confirme que la rupture est principalement due aux anciennetés élevées

Total sinistres



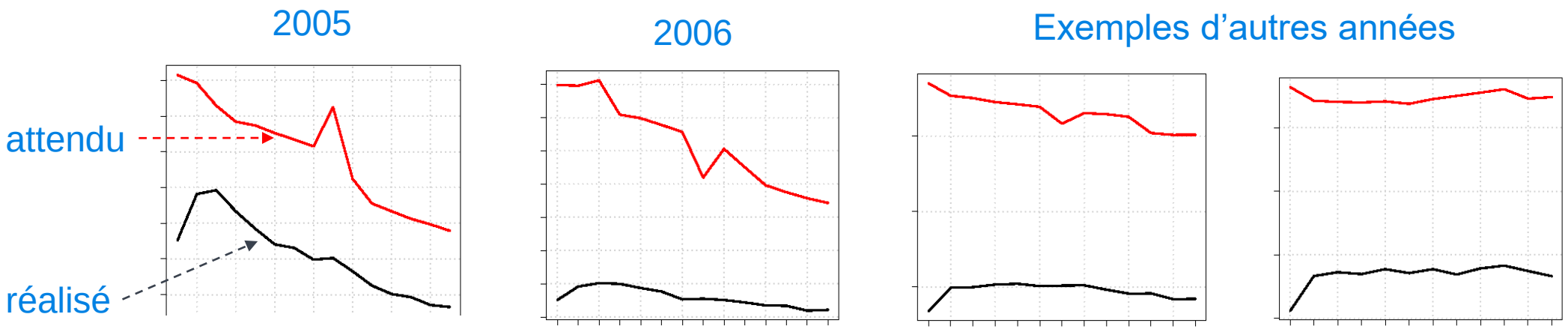
Zoom sur les anciennetés élevées



Applications de l'isolation forest

Outliers de Type 2 – Tendance de sinistres attendus

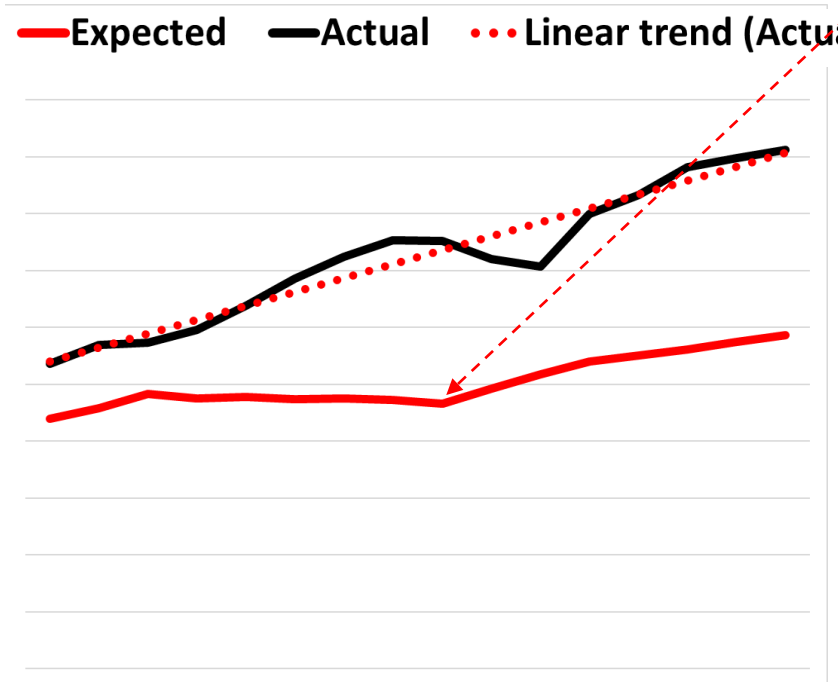
- Les observations correspondant à des souscriptions en 2005 et 2006 et au segment « Sickness hospitalization » sont détectées comme des outliers par l'algorithme
- Les graphiques des sinistres attendus sont relativement stables pour toutes les années de souscription, excepté les années 2005 et 2006

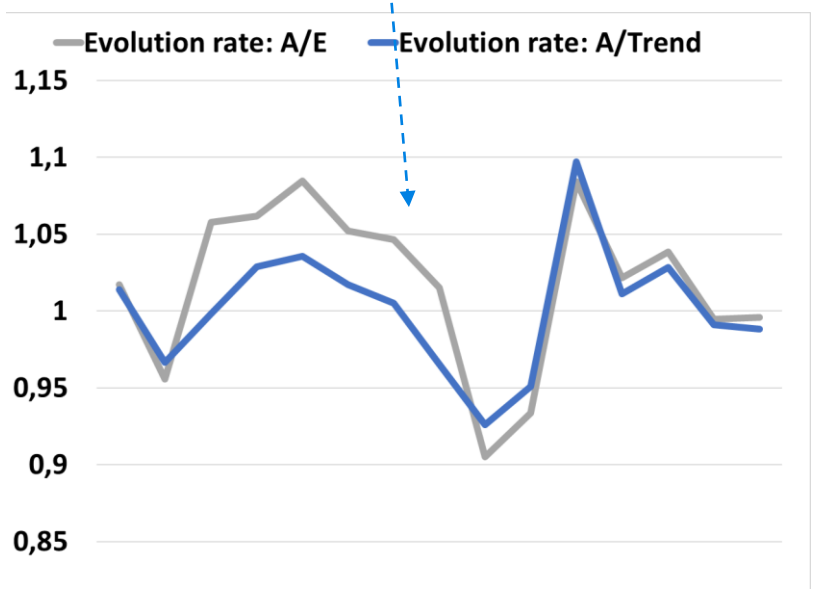


Applications de l'isolation forest

Outliers de Type 3 – Dépendance à la date d'observation

- Segment « Surgery » : la plupart des observations correspondant sont détectées comme des outliers
- Les sinistres attendus (expected) présentent une rupture de tendance à une date de l'historique
- Cette rupture induit des variations sur les taux d'évolution; à titre d'illustration, on compare les taux d'évolution à ceux obtenus sur la base de sinistres attendus en phase avec la tendance des sinistres réalisés ; ces nouveaux taux d'évolution présentent des variations moins importantes





Applications de l'isolation forest

Un exemple d'application à la sinistralité non-vie

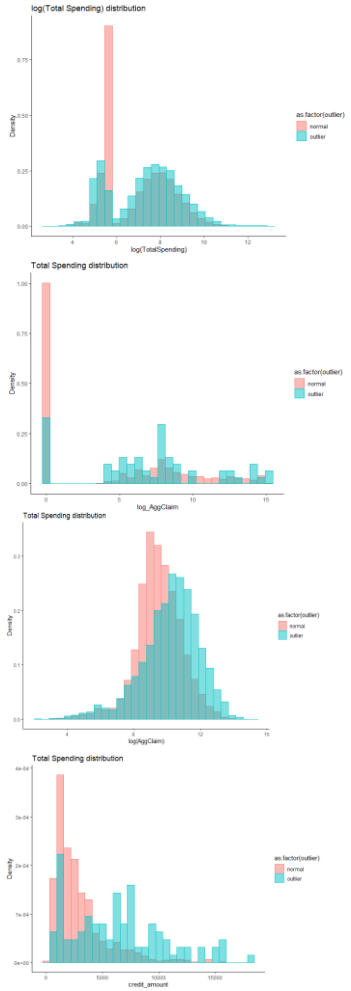
- Représentation de matrices de confusion, comparant le caractère 'outlier' des sinistres avec leur coût

Database 1	Datasets mix	
	Normal	Outlier
Attritional	173540	19761
Large	1453	500
Part of Large in each predicted sub-dataset	**0,83%**	**2,47%**

Database 2	Datasets mix	
	Normal	Outlier
Attritional	512	56
Large	4	2
Part of Large in each predicted sub-dataset	**0,78%**	**3,45%**

Database ausautoBI8999	Datasets mix	
	Normal	Outlier
Attritional	19694	2131
Large	148	73
Part of Large in each predicted sub-dataset	**0,75%**	**3,31%**

Database credit	Datasets mix	
	Normal	Outlier
Attritional	897	93
Large	3	7
Part of Large in each subset	**0,33%**	**7,00%**



Normal set

Outlier set

- La composition de la base des 'outliers' diffère significativement en termes de représentativité des sinistres graves
- Une plus grande proportion de sinistres graves est en effet observée parmi les scores les plus importants

Disclaimer

This presentation presents information of a general nature. It is not intended to guide or determine any specific individual situation and Milliman recommends that users of this presentation will seek explanation and/or amplification of any part of the presentation that they consider not to be clear. Neither the presenter nor the presenter's employer shall have any responsibility or liability to any person or entity with respect to damages alleged to have been caused directly or indirectly by the content of this presentation. All persons who choose to rely in any way on the contents of this presentation do so entirely at their own risk.

The contents of this presentation are confidential and must not be modified, copied, quoted, distributed or shown to any other parties without Milliman's prior written consent.