Data Science

M2 Actuariat

Session 1 : Apprentissage statistique / automatique et actuariat

François HU

Responsable du pôle Intelligence Artificielle, Milliman R&D

Enseignant ISFA

2025

ISFA
Institut de Science Financière et d'Assurances
GRANDE ÉCOLE D'ACTUARIAT & GESTION DES RISQUES

Milliman

Sommaire du cours Data Science

Apprentissage statistique et lien avec l’Actuariat

- **Le « Pourquoi » :** D'un monde de peu de données à un océan d'informations
- **Le « Connu » :** Rappel sur les statistiques inférentielles
- **Le « Nouveau » :** Introduction à l'apprentissage statistique
- **Inférence vs. Prédiction :** Deux objectifs, deux philosophies, une complémentarité essentielle pour l'actuaire.
- **Les Grandes Familles du Machine Learning :** Une cartographie pour s'orienter.

1. Des statistiques à l'apprentissage statistique

- **Le changement de paradigme :** contexte et motivations
- Deux objectifs fondamentaux : Inférence vs. Prédiction
- Apprentissage statistique en pratique : cartographie et applications

ISFA
Institut de Science Financière et d'Assurances

Evolution des données en assurance : du manque à l'abondance

	Avant 2010 : le monde de la rareté	Après 2010 : le monde de l'abondance
Source	Questionnaires, déclarations manuelles.	Objets connectés (IoT), télématique, textes, images, interactions web.
Nature	Données **structurées** (tableaux nets).	Données **hétérogènes** (majoritairement non structurées).
Volume	Faible (kilooctets, mégaoctets).	Massif (pétaoctets).
Fréquence	Statique, collecte annuelle.	Dynamique, flux en temps réel (streaming).
Coût	Élevé à la collecte, faible au stockage.	Faible à la collecte, élevé au traitement.

Evolution des données en assurance : du manque à l'abondance

	Avant 2010 : le monde de la rareté	Après 2010 : le monde de l'abondance
Source		Objets connectés (IoT), télématique, textes, images, interactions web.
Nature		Données **hétérogènes** (majoritairement non structurées).
Volume		Massif (pétaoctets).
Fréquence		Dynamique, flux en temps réel (streaming).
Coût	Élevé à la collecte, faible au stockage.	Faible à la collecte, élevé au traitement.

Caractérisation du Big Data (règle des 3V/4V)

- En grand **Volume** (énorme base de données)
- En grande **Variété** (Numérique, textes, images, vidéos, …)
- A grande **Vitesse** (fréquence d'arrivée de l'information, évolution des données, …)
- (Création de **Valeur** par l'exploitation de ces données)

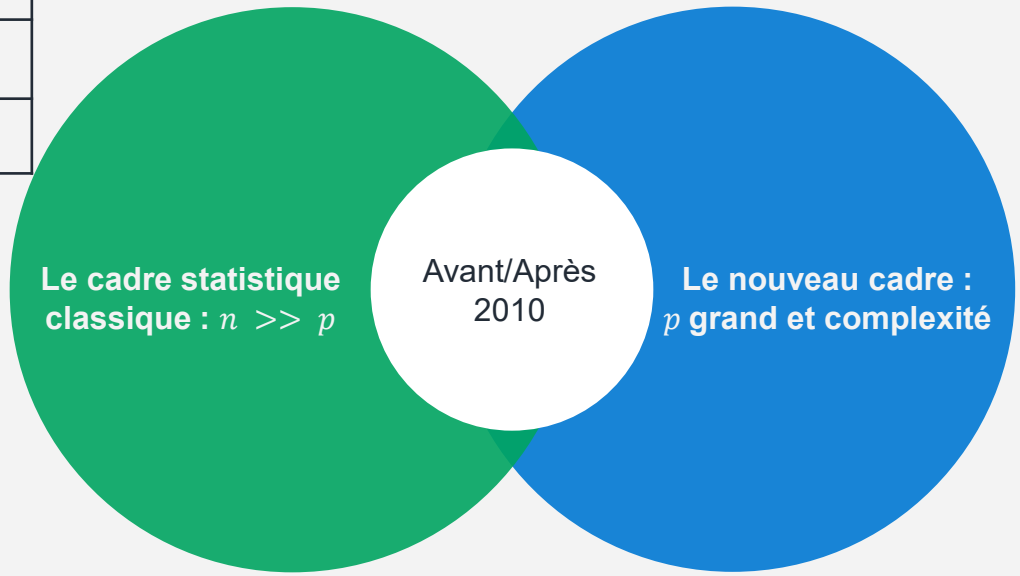
Impact sur la modélisation : le défi de la complexité

p variables explicatives (*Features*) + une variable cible

n observations / échantillon (*sample*)

	F_1	…	F_p	Y
X_1				
…				
X_n				

- **Formalisme :** Le nombre d'observations n est très supérieur au nombre de variables p.
- **Hypothèses implicites :** Relations simples (linéarité, log-linéarité), peu d'interactions, spécifiées manuellement par l'expert.
- **Conséquences :** Les estimateurs (ex: maximum de vraisemblance) sont stables, non biaisés, et leurs propriétés asymptotiques sont bien connues. L'inférence est robuste.



- **Formalisme :** Le nombre de variables p peut être du même ordre de grandeur que n, voire $p > n$. C'est le phénomène de haute dimensionnalité.
- **Défis induits :**
  - **Le fléau de la dimensionnalité ("Curse of Dimensionality") :** L'espace des variables devient immense et creux, les données deviennent éparses, rendant l'estimation difficile.
  - **Instabilité des estimateurs :** En p > n, la solution du maximum de vraisemblance pour un modèle linéaire n'est plus unique. La variance des estimateurs explose.
  - **Sur-apprentissage (Overfitting) :** Avec trop de variables, le modèle a une flexibilité telle qu'il peut "mémoriser" les données d'entraînement, y compris le bruit. Sa capacité à généraliser sur de nouvelles données s'effondre.
  - **Découverte d'interactions :** Il est humainement impossible de tester toutes les interactions pertinentes entre des centaines de variables.

ISFA Institut de Science Financière et d'Assurances

Une nouvelle ère pour la modélisation actuarielle

1. Le facteur limitant a changé.

La contrainte n'est plus la *disponibilité* des données, mais notre *capacité* à en extraire de l'information pertinente.

2. La complexité est devenue la norme.

Les modèles actuariels doivent désormais pouvoir gérer nativement la haute dimensionnalité ($p > n$), les relations non-linéaires et la détection automatique d'interactions.

3. Un besoin méthodologique clair est apparu.

L'apprentissage statistique n'est pas une alternative qui remplace les statistiques, mais une **extension nécessaire** de la boîte à outils de l'actuaire, spécifiquement conçue pour relever ces nouveaux défis.

1. Des statistiques à l'apprentissage statistique

Le changement de paradigme : contexte et motivations

▶ **Deux objectifs fondamentaux** : Inférence vs. Prédiction

Apprentissage statistique en pratique : cartographie et applications

ISFA
Institut de Science Financière et d'Assurances

Le paradigme de l'inférence : expliquer et valider

Quelle est la nature et la force de la relation entre X et Y ?

Approche de modélisation :

On postule un modèle paramétrique a priori. On fait une hypothèse forte sur la forme de la relation.

- *Formalisme :* $g(E[Y|X]) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$. On suppose que la relation est (log-)linéaire et additive.

L'objectif est d'estimer les paramètres β_j**.** Ces coefficients ont une interprétation métier directe.

- *Exemple :* $\beta_{\hat{a}ge}$ représente l'augmentation du log de la prime pour chaque année d'âge supplémentaire, *toutes choses égales par ailleurs*.

On quantifie l'incertitude de ces estimations.

- *Outils :* Erreurs standards, intervalles de confiance, tests d'hypothèses (e.g., H_0: $\beta_j = 0$). La p-valeur est la star de ce paradigme.

Le paradigme de la prédiction : estimer et généraliser

Étant donné un nouveau X, quelle est la meilleure estimation possible pour Y ?

Approche de modélisation :

On traite la fonction f comme une « boîte noire ». On fait peu d'hypothèses sur sa forme.

- *Formalisme :* On suppose seulement l'existence de la relation $Y = f(X) + \varepsilon$, où f peut être arbitrairement complexe et non-linéaire.

L'objectif est de trouver un estimateur $\hat{f}$ qui minimise l'erreur de prédiction. On se soucie peu de la structure interne de $\hat{f}$.

- *Exemple :* Un réseau de neurones avec des milliers de paramètres. Il est impossible d'interpréter chaque paramètre individuellement.

On évalue la performance sur des données jamais vues.

- *Outils :* L'erreur de généralisation, estimée via un jeu de test ou par validation croisée. Les métriques comme la RMSE (Root Mean Squared Error) ou l'AUC (Area Under the Curve) sont les stars de ce paradigme.

Une dichotomie essentielle pour l'actuaire

Critère	Inférence Statistique (Comprendre)	Apprentissage Statistique (Prédire)
Objectif Principal	Estimer des paramètres (β) et quantifier leur incertitude.	Trouver une fonction ($\hat{f}$) qui minimise l'erreur de prédiction.
Complexité du Modèle	Les modèles **simples et interprétables** sont préférés.	Les **modèles complexes sont acceptés** s'ils améliorent la performance.
Évaluation	Critères "internes" au modèle : R^2, p-valeurs, tests d'adéquation (AIC/BIC).	Critères "externes" : performance sur un jeu de test (RMSE, AUC, etc.).
Le Modèle est...	... une description plausible du processus générateur des données.	... un algorithme qui mappe efficacement les entrées aux sorties.
Dilemme Actuariel	**Transparence & Justification :** Comment justifier un tarif au régulateur ?	**Performance & Compétitivité :** Comment avoir le tarif le plus précis du marché ?

Une dichotomie essentielle pour l'actuaire

1. L'inférence se concentre sur les paramètres d'un modèle postulé.

Son but est l'explication et la validation d'hypothèses.

2. La prédiction se concentre sur la performance d'une fonction estimée.

Son but est la généralisation à de nouvelles données.

3. La distinction guide tout le processus de modélisation.

Le choix de l'algorithme, la gestion de sa complexité et, surtout, la méthode d'évaluation en dépendent directement.

4. Le rôle de l'actuaire est de maîtriser et de réconcilier ces deux objectifs.

1. Des statistiques à l'apprentissage statistique

Le changement de paradigme : contexte et motivations

Deux objectifs fondamentaux : Inférence vs. Prédiction

- **Apprentissage statistique en pratique** : cartographie et applications

ISFA
Institut de Science Financière et d'Assurances

Taxonomie des méthodes d'apprentissage

Une classification par la nature de l'information

	Apprentissage Supervisé	Apprentissage Non Supervisé	Apprentissage par Renforcement
Information disponible	On dispose d'exemples (X, Y) où la « bonne réponse » (le label / l'étiquette) Y est connue.	On ne dispose que des entrées X, sans étiquette ou réponse associée.	Un agent interagit avec un environnement et reçoit des signaux (récompenses/punitions).
Question clé	Comment prédire Y à partir de X ?	Quelle est la structure cachée dans les données X ?	Quelle est la meilleure séquence d'actions à prendre ?
Usage en assurance	**Largement dominant.** La majorité des cas d'usage (tarification, fraude, churn...)	**Spécifique et complémentaire** (segmentation de portefeuille, détection d'anomalies).	**Émergent** et avancé (gestion ALM, tarification dynamique).

Une enquête de Kaggle de 2021 auprès des data scientists (tous secteurs confondus) montrait que les algorithmes de régression et de classification (supervisés) étaient utilisés par plus de **80% des répondants**, loin devant le **clustering (~30%) (non supervisé).**
Source : Kaggle, "2021 State of Machine Learning and Data Science".

Cette proportion est généralement considérée comme encore plus élevée dans le secteur de l'assurance qui est très orienté vers la prédiction de cibles métier précises.

Les applications de l'apprentissage par renforcement en assurance **sont encore majoritairement au stade de la recherche et développement**. Voir par exemple : Wüthrich, M. V. (2020). *The Actuary in the 21st Century: Reinforcement Learning*. SSRN.

ISFA
Institut de Science Financière et d'Assurances

Zoom sur l'apprentissage supervisé : le cœur de la modélisation actuarielle

Au sein de **l'apprentissage supervisé**, la tâche est définie par **la nature de la variable cible Y** que l'on cherche à prédire.

1. La Régression : Prédire une quantité continue

– Formellement : La variable cible Y est quantitative ($Y \in \mathbb{R}$).

– Questions actuarielles :

- *Quel sera le **montant** du prochain sinistre corporel ?*
- *Quelle sera la **charge de sinistre totale** pour ce contrat auto l'an prochain ?*
- *Quel est le **nombre de jours** d'arrêt de travail attendu ?*

2. La Classification : Prédire une catégorie discrète

– Formellement : La variable cible Y est qualitative ($Y \in \{c_1, \dots, c_K\}$).

– Questions actuarielles :

- *Cette déclaration de sinistre est-elle **frauduleuse ou légitime** ? (Classification binaire, K=2)*
- *Ce client va-t-il **résilier son contrat** dans les 3 prochains mois ? (Classification binaire)*
- *Ce sinistre concerne-t-il un **bris de glace, un vol, ou un accident responsable** ? (Classification multi-classes, K=3)*

Exercice interactif : régression ou classification ?

Instructions *:* Pour chaque problème, déterminez s'il s'agit d'une tâche de régression ou de classification.

Estimer le coût final d'un sinistre IARD en cours de gestion

Identifier les clients susceptibles de répondre positivement à une campagne marketing

Prédire la durée de survie d'un nouvel assuré en assurance-vie

Orienter automatiquement un email entrant vers le bon service de gestion (Sinistres, Souscription, Comptabilité)

Attribuer un score de risque de 1 à 100 à un nouveau contrat

Régression

Classification

De nouvelles perspectives pour le métier d'actuaire

Apprentissage statistique n'est pas qu'un outil technique, c'est un levier stratégique qui transforme les missions de l'actuaire

Tarification : de la segmentation à l'hyper-personnalisation

Avant : Quelques grands segments de risque (GLM)

Maintenant : Tarification comportementale (télématique), modélisation d'interactions fines, pricing individualisé

Détection de fraude : de l'expertise manuelle au scoring systématique

Avant : Règles métier, audits a posteriori

Maintenant : Score de suspicion en temps réel sur 100% des sinistres, détection de réseaux de fraudeurs (non supervisé)

Rétention client (Churn) : du réactif au proactif

Avant : Analyser a posteriori pourquoi les clients sont partis

Maintenant : Prédire *qui* va partir pour lancer des actions de rétention ciblées et rentables

L'écosystème d'outils de l'actuaire-data scientist

La mise en œuvre de ces modèles repose sur un écosystème logiciel open-source, mature et puissant

Langages de programmation :

- **Python :** Le standard de l'industrie pour la data science et le machine learning. Vaste écosystème de bibliothèques.
- **R :** Le langage historique des statisticiens, extrêmement puissant pour l'analyse de données et la modélisation statistique.

Bibliothèques incontournables :

- **Python : Pandas** (manipulation de données), **Scikit-learn** (algorithmes de ML), **Matplotlib/Seaborn** (visualisation).
- **R** *:* L'écosystème **Tidyverse** (manipulation/visualisation), **Caret** ou **Tidymodels** (workflow de modélisation).

Environnement de travail :

- Les **Notebooks** (Jupyter, R Markdown) sont devenus le standard pour la recherche et le prototypage, car ils permettent de mêler code, résultats, visualisations et texte.

2. Le processus de modélisation et les concepts principaux

- **Processus de modélisation et le feature engineering**
- Compromis Biais-Variance : la théorie de l'apprentissage
- Évaluation robuste des modèles : de la théorie à la pratique

ISFA
Institut de Science Financière et d'Assurances

Le mythe et la réalité du travail d'un data scientist

Une méthodologie standard : le cycle de vie CRISP-DM

CRISP-DM : Cross Industry Standard Process for Data Mining

Pour structurer ce processus, une méthodologie a été développée et reste la référence dans l'industrie : **CRISP-DM.**

Ce n'est pas un processus linéaire (en cascade), mais un **cycle itératif**.

Les 6 phases :

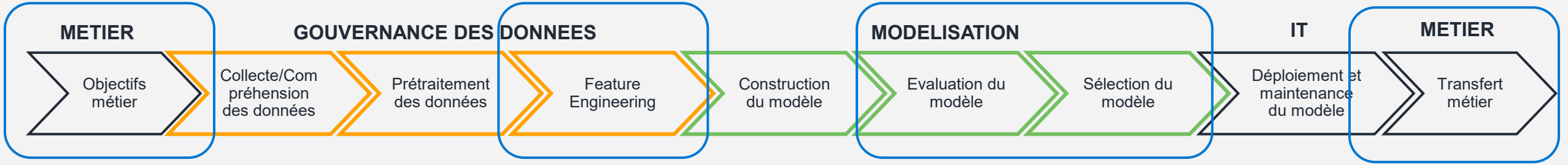
1. **Business Understanding** (Compréhension du besoin métier)
2. **Data Understanding** (Compréhension des données)
3. **Data Preparation** (Préparation des données)
4. **Modeling** (Modélisation)
5. **Evaluation** (Évaluation)
6. **Deployment** (Déploiement)

ISFA
Institut de Science Financière et d'Assurances

Le cycle de vie d'un projet data science

Une autre perspective

Dans cette session, parlons de tout sauf les techniques de modélisation !



Le besoin des assureurs

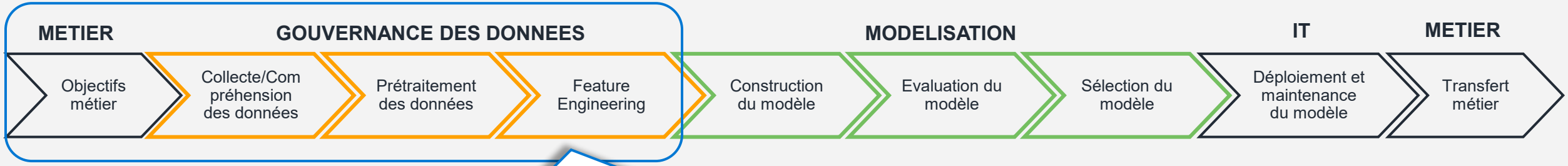
L'ingénierie des données doit être un sujet central en Data Science !

Validation du modèle d'apprentissage

Les phases amont : définir le problème et préparer le terrain

Une autre perspective

Dans cette session, parlons de tout sauf les techniques de modélisation !

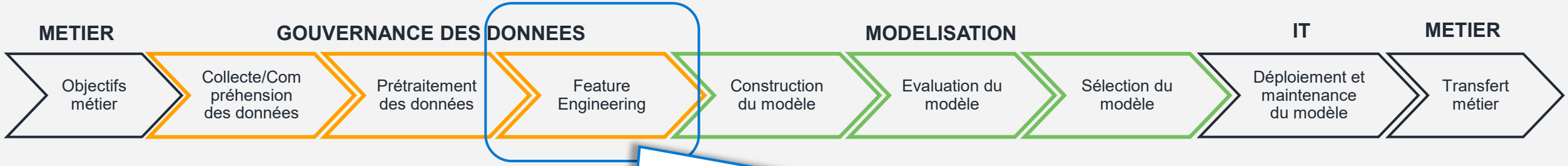


1. **Compréhension du besoin métier**
   Traduire un objectif business en un problème de modélisation bien défini (régression/classification) avec une métrique de succès claire.
2. **Compréhension des données**
   Analyse exploratoire (EDA) pour comprendre la structure, la qualité et les limites de vos données brutes.
3. **Préparation des données**
   Cette phase inclut le nettoyage (valeurs manquantes, outliers...) et une étape si fondamentale qu'elle mérite son propre focus : la création de variables, ou ***feature engineering***.

Zoom sur le feature engineering

le secret des meilleurs modèles

Dans cette session, parlons de tout sauf les techniques de modélisation !



Le feature engineering est le processus qui consiste à utiliser la connaissance du domaine pour transformer des données brutes en variables (features) qui représentent mieux le problème et améliorent la performance des modèles prédictifs.

Pourquoi est-ce si important ?

- Les modèles d'apprentissage, même les plus complexes, sont limités par les données qu'on leur fournit.
- De bonnes variables permettent à des modèles simples de devenir très performants.
- C'est l'étape où la connaissance actuarielle (compréhension du risque, de la réglementation, du comportement client) apporte le plus de valeur.

Une citation célèbre :

"Coming up with features is difficult, time-consuming, requires expert knowledge. 'Applied machine learning' is basically feature engineering." - Andrew Ng (Professeur à Stanford, pionnier de l'IA).

Le feature engineering en pratique actuarielle

Comment obtenir de meilleurs modèles ?

Combinaisons et ratios

- **Ratio** Sinistres / Primes : Mesure de la sinistralité passée d'un client.
- **Ratio** montant_financé / valeur_véhicule : Indicateur de risque moral en assurance auto.

Manipulations de dates et durées

- **Âge** du conducteur, **Ancienneté** du permis.
- **Ancienneté du contrat** : Un indicateur clé de la fidélité.
- **Jour de la semaine de l'accident** : Les accidents du vendredi soir sont-ils plus graves ?
- **Mois de la déclaration** : Pour capturer des effets de saisonnalité (ex: gel en hiver).

Discrétisation (Binning)

- **Transformer** une variable continue en catégories pour capturer des effets non-linéaires.

Exemple : Âge en [18-25], [26-40], [41-65], [66+]. Très utile pour les GLM.

Traitement des variables catégorielles à haute cardinalité

Problème : Que faire d'une variable comme le code postal ou la profession avec des centaines de modalités ?

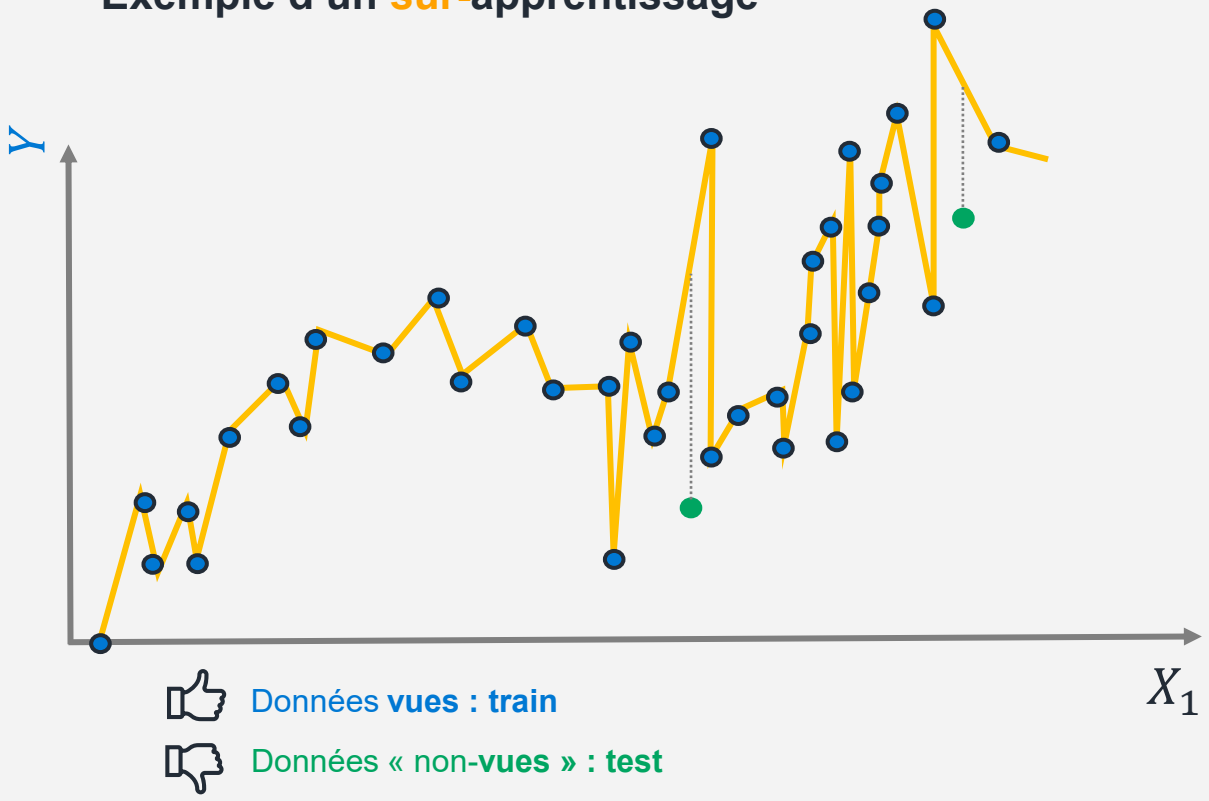
Solutions : Agrégation (ex: regrouper par département), création de variables basées sur la fréquence ou le risque moyen de la catégorie (Target Encoding).

En aval de la modélisation : le piège du sur-apprentissage (1/3)

Sur-apprentissage et sous-apprentissage

⚠ **Les choix de modélisation doivent bien se généraliser aux données « non vues ».**

Exemple d'un sur-apprentissage



- **Définition intuitive :** Un modèle est en sur-apprentissage lorsqu'il apprend "par cœur" les données d'entraînement, y compris leurs particularités aléatoires (le "bruit").
- **Conséquence :** Le modèle devient excellent pour prédire les données qu'il a déjà vues, mais très mauvais pour prédire de nouvelles données. Il a perdu sa capacité de **généralisation**.
- **Analogie :** L'étudiant qui mémorise les solutions des annales d'examen sans comprendre les concepts. Il aura 20/20 si le même examen retombe, mais 0/20 sur un nouvel énoncé.
- **Cause principale :** Le modèle est trop complexe ou trop flexible par rapport à la quantité d'information réellement présente dans les données.

ISFA Institut de Science Financière et d'Assurances

Exemple illustratif

Le piège du sur-apprentissage (2/3)

Sur-apprentissage et sous-apprentissage

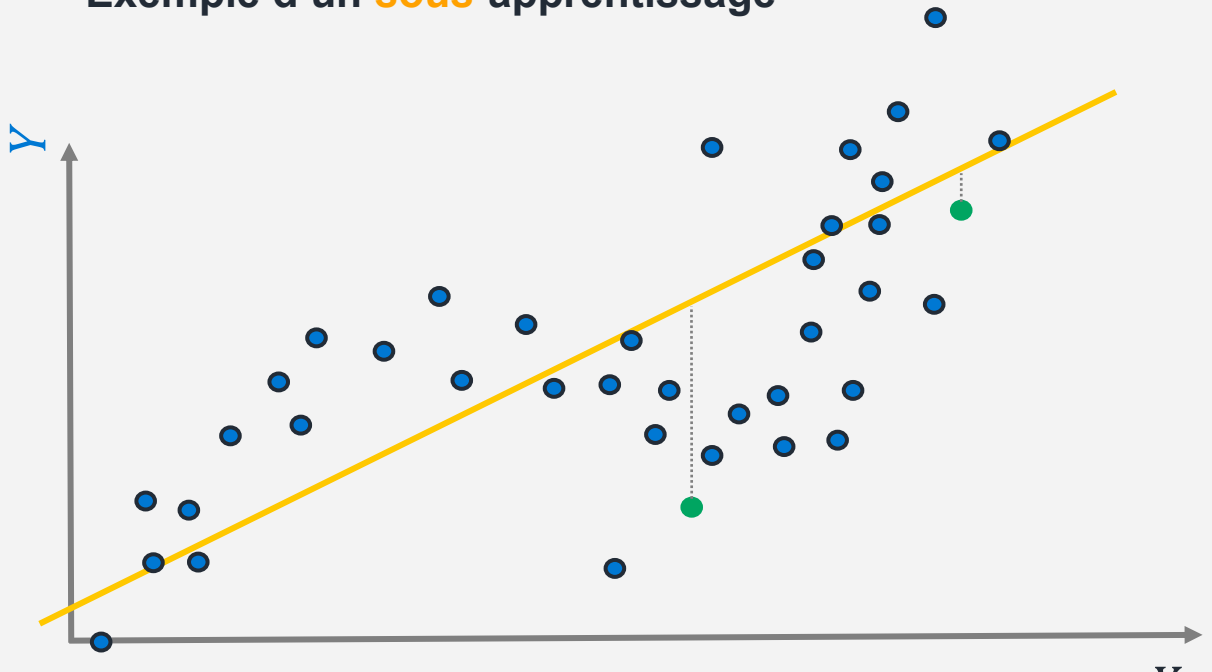
⚠ **Les choix de modélisation doivent bien se généraliser aux données « non vues ».**

Exemple d'un sur-apprentissage

Données **vues : train**

Données « non-**vues** » : **test**

Exemple d'un sous-apprentissage



Données **vues : train**

Données « non-**vues** » : **test**

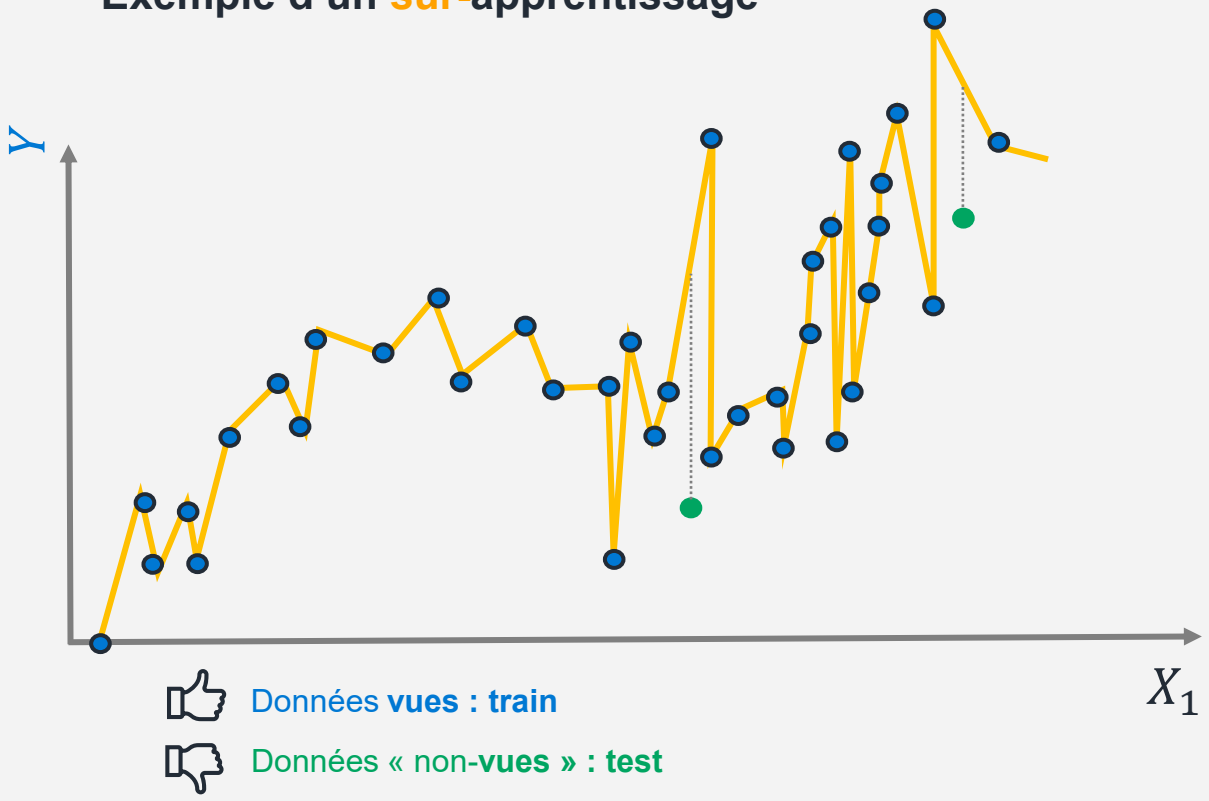
Exemple illustratif

Le piège du sur-apprentissage (3/3)

Sur-apprentissage et sous-apprentissage

⚠ **Les choix de modélisation doivent bien se généraliser aux données « non vues ».**

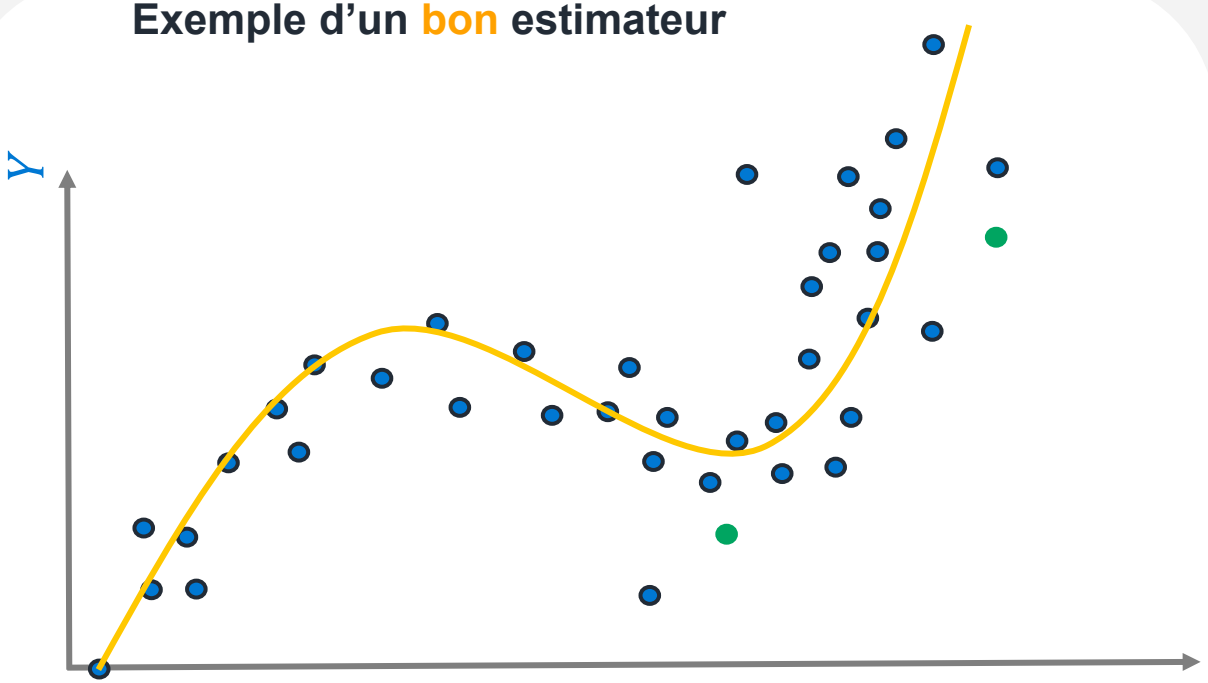
Exemple d'un sur-apprentissage



Données **vues : train**

Données « non-**vues** » : **test**

Exemple d'un bon estimateur



Données **vues : train**

Données « non-**vues** » : **test**

Exemple illustratif

Les conséquences actuarielles du sur-apprentissage

Le s**ur-apprentissage** est un **risque financier** et **opérationnel** concret.

En Tarification :

Un modèle sur-appris créera des tarifs très instables, réagissant de manière excessive à des caractéristiques non pertinentes.

Il identifiera de fausses "niches de bons risques" qui n'existent pas, menant à des politiques de souscription basées sur du bruit.

Risque : Anti-sélection massive. Vous attirez des risques que vous pensez avoir bien tarifés, mais qui sont en réalité sous-tarifés.

En Provisionnement :

Un modèle de coût ultime sur-appris sera incapable de prédire correctement l'évolution des sinistres futurs, car il sera trop collé aux fluctuations passées.

Risque : Sous-estimation ou sur-estimation systématique des réserves, avec un impact direct sur le bilan et la solvabilité.

En Détection de Fraude :

Un modèle sur-appris identifiera des schémas de fraude extrêmement spécifiques aux données passées et ratera toutes les nouvelles formes de fraude.

Risque : Inefficacité opérationnelle. Les gestionnaires perdent confiance dans l'outil qui génère trop de fausses alertes ou manque les vraies fraudes.

2. Le processus de modélisation et les concepts principaux

Processus de modélisation et le feature engineering

▶ **Compromis Biais-Variance :** la théorie de l'apprentissage

Évaluation robuste des modèles : de la théorie à la pratique

ISFA
Institut de Science Financière et d'Assurances

Formalisation de notre objectif : minimiser l'erreur de généralisation

Le modèle statistique :

Nous postulons l'existence d'une relation

$$Y = f(X) + \varepsilon \quad \text{où} \quad E[\varepsilon] = 0 \text{ et } Var(\varepsilon) = \sigma^2.$$

X et Y sont des variables aléatoires. f est une **fonction déterministe inconnue**. ε est le **bruit**, l'aléa ou **l'erreur irréductible** indépendante de X.

Le processus d'apprentissage :

Nous observons un **échantillon d'entraînement** $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$, considéré comme une réalisation de (X, Y).

À partir de D, nous construisons un estimateur de f, noté $\hat{f}$.

L'objectif :

Nous voulons que $\hat{f}$ **soit un bon prédicteur pour une nouvelle observation** (X_0, Y_0) indépendante de D.

La mesure de l'erreur (Loss Function) :

Nous utilisons la **perte quadratique** :

$$L(Y_0, \hat{f}(X_0)) = (Y_0 - \hat{f}(X_0))^2$$

L'erreur de généralisation :

C'est l'erreur quadratique moyenne de $\hat{f}$ sur une nouvelle observation (x_0, y_0).

$$\mathrm{Err}(X_0) = \mathrm{E}[(Y_0 - \hat{f}(X_0))^2]$$

La décomposition de l'erreur : biais, variance et erreur irréductible

On part de $Err(x_0) = E[(y_0 - \hat{f}(x_0))^2]$.

Isoler le bruit irréductible ε

En utilisant $y_0 = f(x_0) + \varepsilon$, l'erreur devient :

$$Err(x_0) = E[(f(x_0) - \hat{f}(x_0))^2] + E[\varepsilon^2]$$

$$Err(x_0) = E[(f(x_0) - \hat{f}(x_0))^2] + \sigma^2$$

Décomposer l'erreur du modèle

On décompose le premier terme en ajoutant et soustrayant $E[\hat{f}(x_0)]$:

$$E[(f(x_0) - \hat{f}(x_0))^2] = \underbrace{(f(x_0) - E[\hat{f}(x_0)])^2}_{(Biais[\hat{f}(x_0)])^2} + \underbrace{E[(\hat{f}(x_0) - E[\hat{f}(x_0)])^2]}_{Variance[\hat{f}(x_0)]}$$

$$Err(x_0) = (Biais[\hat{f}(x_0)])^2 + Variance[\hat{f}(x_0)] + \sigma^2$$

Zoom sur le biais : l'erreur de simplification

Définition formelle :

$$Biais[\hat{f}\,(x_0)] = E[\hat{f}\,(x_0)] - f(x_0)$$

Interprétation : C'est la différence entre la prédiction *moyenne* de notre modèle et la vraie valeur. Un biais élevé signifie que notre modèle, en moyenne, rate systématiquement la cible.

Causes : Le modèle choisi est trop simple ou fait des hypothèses trop fortes sur la nature de f.

Exemple actuariel : Utiliser un modèle linéaire (GLM) pour tarifer un risque alors que la relation entre l'âge et la sinistralité est fortement non-linéaire (forme en U). Le modèle sera systématiquement faux pour les conducteurs très jeunes et très âgés.

Un biais élevé conduit au **sous-apprentissage (underfitting)**.

Zoom sur la variance : l'erreur d'instabilité

Définition formelle :

$$Variance[\hat{f}\,(x_0)] \;=\; E[(\hat{f}\,(x_0) \;-\; E[\hat{f}\,(x_0)])^2]$$

Interprétation : Mesure à quel point notre modèle $\hat{f}$ changerait si on l'entraînait sur un jeu de données différent. Une variance élevée signifie que le modèle est très instable et sensible aux particularités de l'échantillon.

Causes : Le modèle est trop complexe, trop flexible. Il a trop de « degrés de liberté ».

Exemple actuariel : Entraîner un arbre de décision très profond. L'arbre peut créer des règles très spécifiques comme "SI âge=23 ET voiture=modèle_X ET code_postal=Y...". Cette règle sera peut-être vraie pour les 2 personnes de l'échantillon d'entraînement qui correspondent, mais sera probablement fausse pour toute nouvelle personne.

Une variance élevée conduit au **sur-apprentissage (overfitting)**.

Compromis Biais-Variance : un levier d'optimisation (1/2)

Comprendre les sources d'erreur pour mieux les mitiger

- **Compromis Biais-Variance :**

  Erreur de généralisation (erreur sur données test)

  = **Biais** + **Variance** + **Erreur irréductible**

  Biais : erreur systématique, due à des hypothèses trop simplistes. *Le modèle « rate » la cible*
  → **sous-apprentissage**.

  Variance : erreur liée à la sensibilité aux fluctuations des données d'entraînement. *Le modèle « s'éparpille »*
  → **sur-apprentissage**.

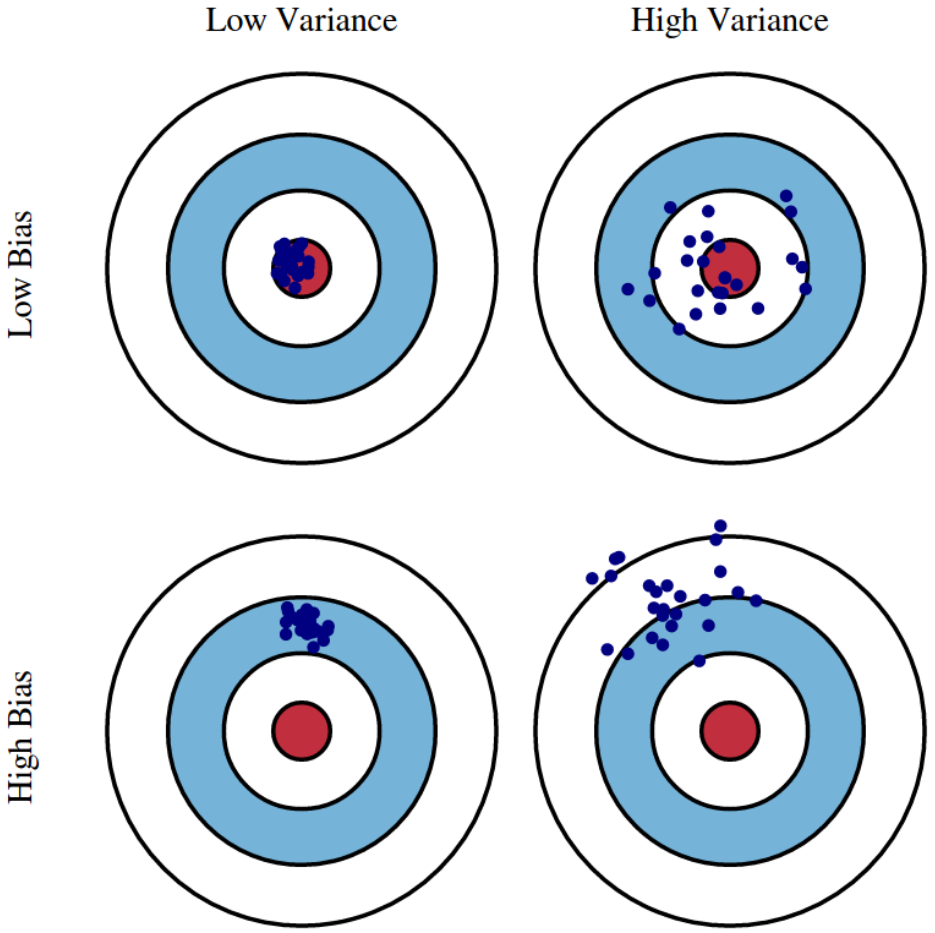
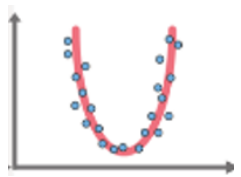


Image Source: http://scott.fortmann-roe.com/docs/BiasVariance.html

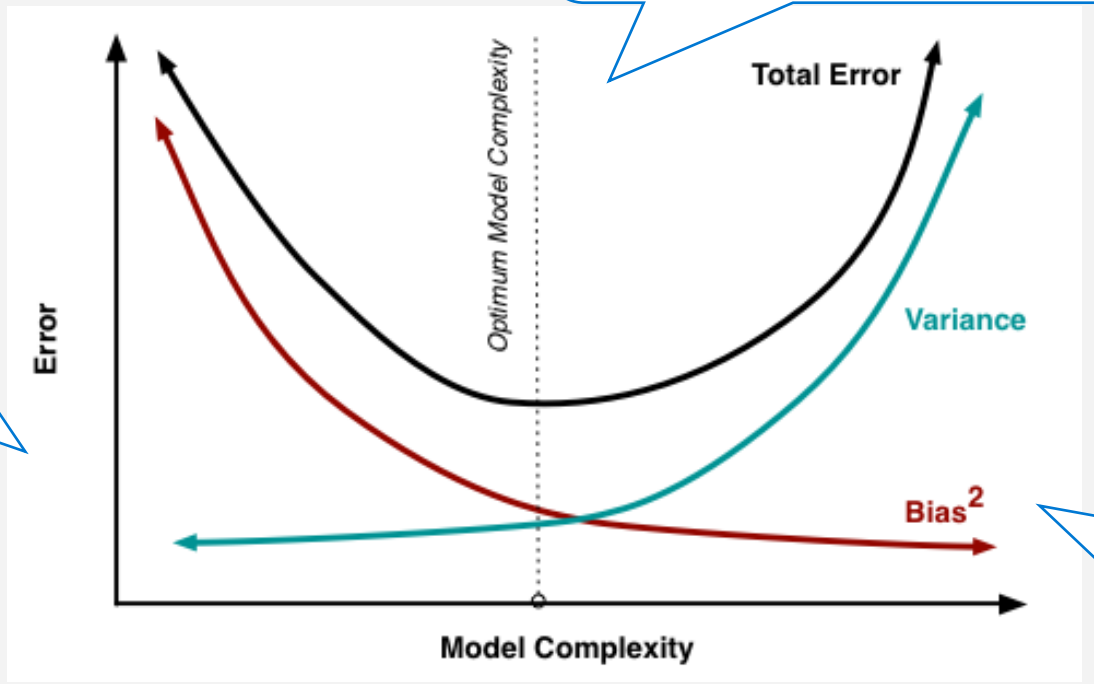
Compromis Biais-Variance : un levier d'optimisation (2/2)

Comprendre les sources d'erreur pour mieux les mitiger

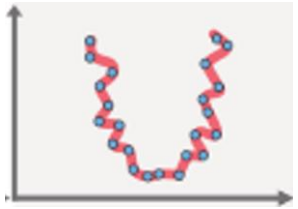
Notre objectif



Sous-apprentissage



Sur-apprentissage



Comment peut-on **estimer l'erreur de généralisation** de notre modèle pour savoir où nous nous situons sur cette courbe ?

Le dilemme de l'actuaire : choisir son compromis

Le compromis biais-variance : une carte des familles de modèles

Le défi n'est pas de trouver "le meilleur modèle" dans l'absolu, mais le modèle qui offre le meilleur compromis **pour le problème et les données considérés**.

Famille de modèles (concept)	Biais typique	Variance typique	Caractéristique principale
Modèles Linéaires (ex: GLM)	Élevé	Faible	Structure rigide, interprétable.
Modèles « Flexibles » (ex: Arbres de décision)	Faible	Élevée	Apprend des règles locales, très adaptable.
Modèles « Agrégés » (ex: Méthodes ensemblistes)	Faible	Moyenne	Combine de nombreux modèles pour être à la fois adaptable et stable.

2. Le processus de modélisation et les concepts principaux

Processus de modélisation et le feature engineering

Compromis Biais-Variance : la théorie de l'apprentissage

- **Évaluation robuste des modèles :** de la théorie à la pratique

ISFA
Institut de Science Financière et d'Assurances

Rappel

Sur-apprentissage et sous-apprentissage

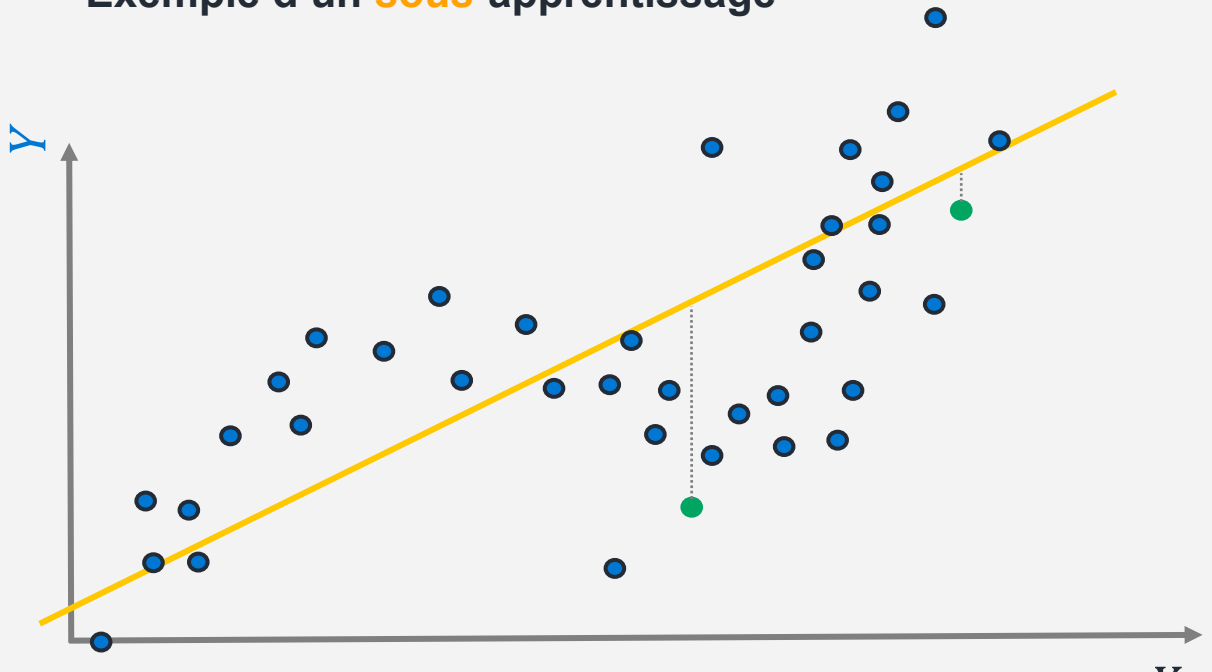
⚠ **Les choix de modélisation doivent bien se généraliser aux données « non vues ».**

Exemple d'un sur-apprentissage

Données **vues : train**

Données « non-**vues** » : **test**

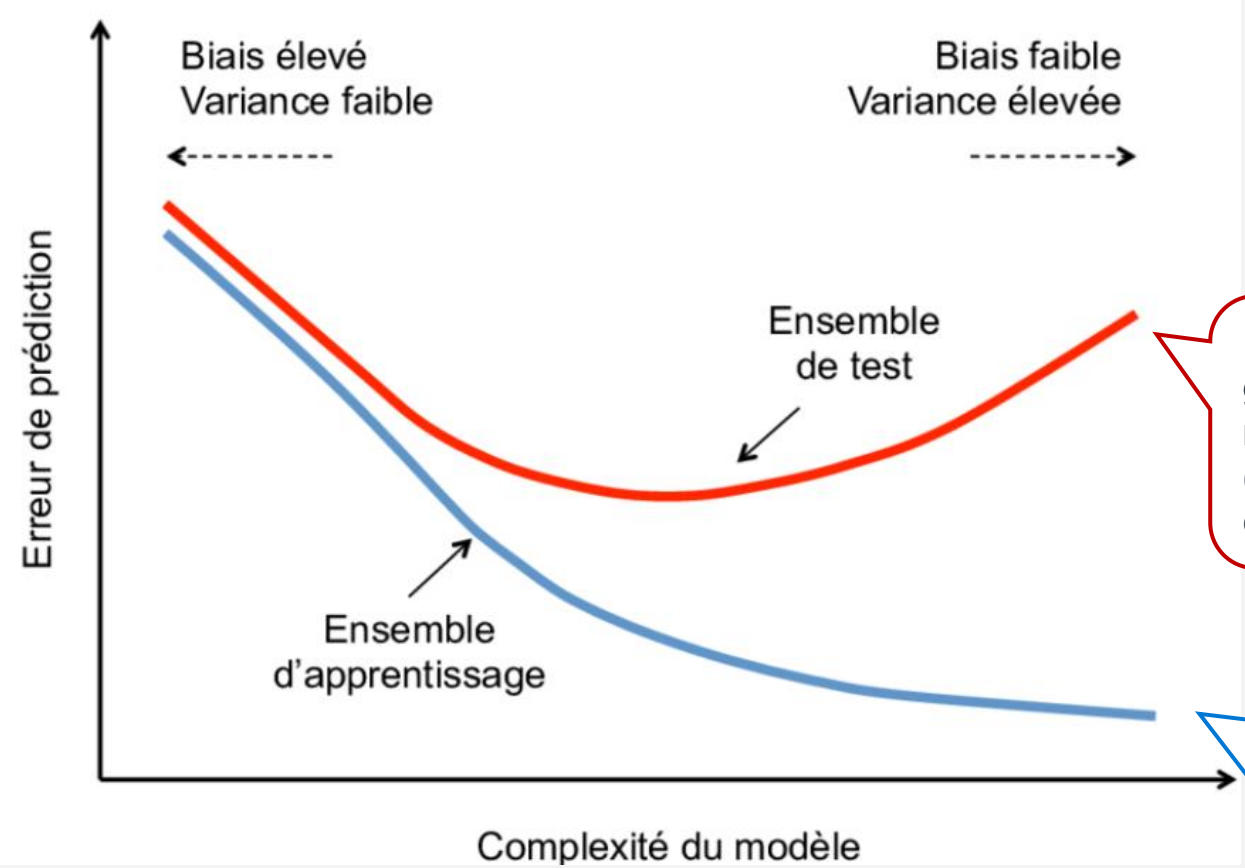
Exemple d'un sous-apprentissage



Données **vues : train**

Données « non-**vues** » : **test**

Le défi : l'erreur d'entraînement est un indicateur trompeur



Pour estimer correctement l'erreur de généralisation, nous devons **évaluer le modèle sur des données nouvelles (ensemble de test)**, indépendantes de celles utilisées pour sa construction.

L'**erreur d'entraînement** est la performance du modèle $\hat{f}$ mesurée sur le jeu de données D qui a servi à l'apprendre.

$$Err\_train = (1/n)\ \Sigma\left(y_i - \hat{f}(x_i)\right)^2$$

pour $(x_i, y_i) \in D$.

La stratégie de base

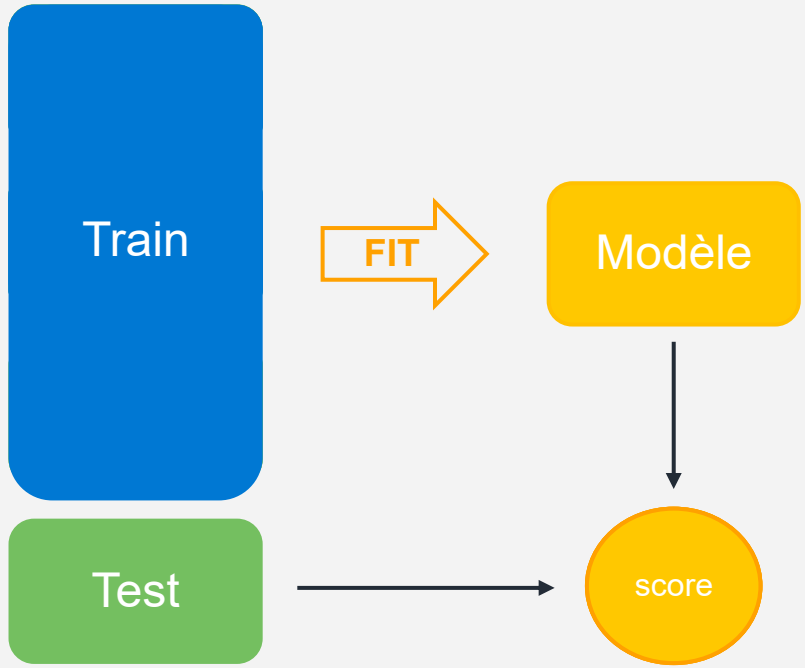
Approche train-test split

⚠ **Les choix de modélisation doivent bien se généraliser aux données de test.**

- **Approche classique :**

Séparer les données en :

1. données d'entraînement (**train**) pour calibrer les coefficients du modèle,
2. (**en option**) données de validation (**valid**) pour calibrer d'autres « hyperparamètres » du modèle
3. données de test (**test**) pour évaluer le modèle



Généralement entre **60% et 80%**
pour le train et 20%-40% pour le test

La validation croisée

Approche « K-folds » : La méthode standard pour obtenir une estimation robuste de la performance.

⚠ **Les choix de modélisation doivent bien se généraliser aux données de test.**

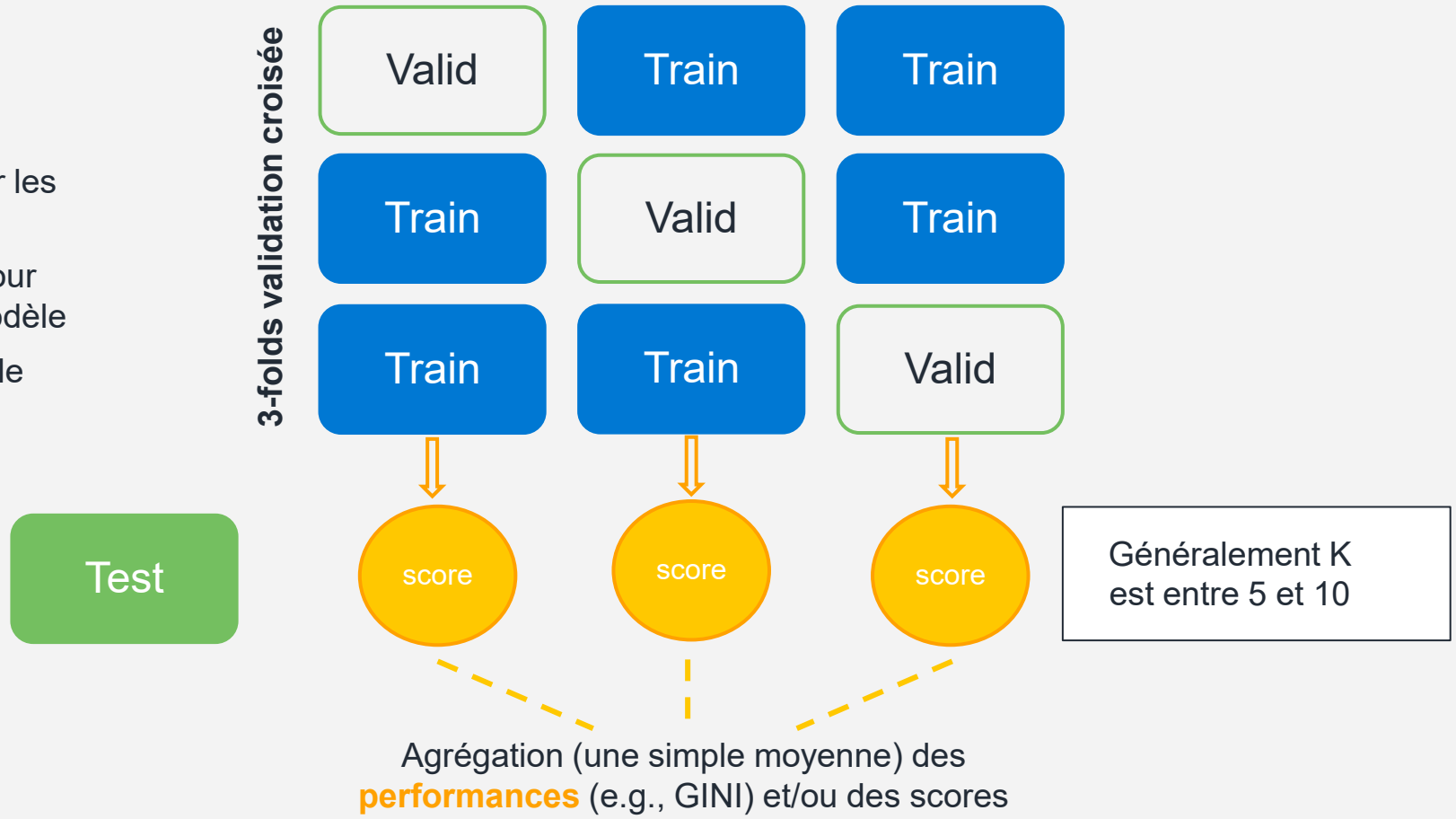
Approche classique :

Séparer les données en :

1. données d'entraînement (**train**) pour calibrer les coefficients du modèle,
2. (**en option**) données de validation (**valid**) pour calibrer d'autres « hyperparamètres » du modèle
3. données de test (**test**) pour évaluer le modèle

▶ **La validation croisée :**

Diviser les données en plusieurs sous-ensembles pour entraîner et tester le modèle **sur K différentes partitions**, ce qui permet d'évaluer la généralisation du modèle sur des données non vues.



ISFA
Institut de Science Financière et d'Assurances

3. Evaluation des modèles et études de cas actuarielles

- **Les métriques d'évaluation :** choisir la bonne règle pour mesurer

Études de cas actuarielles : mettre la théorie en pratique

ISFA
Institut de Science Financière et d'Assurances

Métriques pour les problèmes de régression

Mesurer l'erreur pour les prédictions de quantités continues

Pour la régression, deux métriques dominent : la MAE et la RMSE. **La MAE est robuste, la RMSE est sensible.** La principale différence réside dans la façon dont elles traitent les grosses erreurs. **Le R^2 est plus un outil d'analyse** de la qualité d'ajustement qu'une véritable mesure de la performance prédictive future, surtout en haute dimension.

Erreur Absolue Moyenne (MAE - Mean Absolute Error)

- **MAE**

$$\frac{1}{n} \times \sum_i |y_i - \hat{f}(x_i)|$$

- **Interprétation** : L'erreur moyenne en valeur absolue. Facile à interpréter, dans l'unité de la cible, et moins sensible aux valeurs extrêmes (outliers).

Erreur Quadratique Moyenne (MSE - Mean Squared Error)

- **MSE**

$$\frac{1}{n} \times \sum_i \left(y_i - \hat{f}(x_i)\right)^2$$

- **RMSE (Root MSE)**

$$\sqrt{\frac{1}{n} \times \sum_i \left(y_i - \hat{f}(x_i)\right)^2}$$

- **Interprétation** : La métrique la plus courante. Elle pénalise fortement les grandes erreurs en raison de la mise au carré. Également dans l'unité de la cible.

Coefficient de Détermination (R^2)

$$R^2 = 1 - \frac{\sum(Y_i - \widehat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

Avec Y_i valeur observée, $\widehat{Y}_i$ valeur estimée et $\bar{Y}$ valeur moyenne

- **Interprétation** : La proportion de la variance de Y expliquée par le modèle. Utile pour l'inférence, mais à utiliser avec prudence pour la prédiction, car il augmente mécaniquement avec le nombre de variables.

Le choix MAE vs. MSE : une décision métier

Mesurer l'erreur pour les prédictions de quantités continues

Pour la régression, deux métriques dominent : la MAE et la RMSE. **La MAE est robuste, la RMSE est sensible.** La principale différence réside dans la façon dont elles traitent les grosses erreurs. **Le R^2 est plus un outil d'analyse** de la qualité d'ajustement qu'une véritable mesure de la performance prédictive future, surtout en haute dimension.

Dilemme :

- Un modèle minimisant la MAE peut faire de grandes erreurs occasionnelles, tant que l'erreur moyenne reste faible.
- Un modèle minimisant la RMSE sera contraint d'éviter à tout prix les grandes erreurs, quitte à faire plus de petites erreurs en moyenne.

- **MAE**

$$\frac{1}{n} \times \sum_i \left| y_i - \hat{f}(x_i) \right|$$

- **Interprétation** : L'erreur moyenne en valeur absolue. Facile à interpréter, dans l'unité de la cible, et moins sensible aux valeurs extrêmes (outliers).

- **MSE**

$$\frac{1}{n} \times \sum_i \left(y_i - \hat{f}(x_i) \right)^2$$

- **RMSE (Root MSE)**

$$\sqrt{\frac{1}{n} \times \sum_i \left(y_i - \hat{f}(x_i) \right)^2}$$

- **Interprétation** : La métrique la plus courante. Elle pénalise fortement les grandes erreurs en raison de la mise au carré. Également dans l'unité de la cible.

Question pour l'actuaire (ex: prédiction du coût d'un sinistre) :

- Se tromper un peu sur beaucoup de dossiers ?
- Se tromper énormément sur un très petit nombre de dossiers graves ?

En pratique : En raison de l'aversion au risque et de l'impact disproportionné des sinistres graves, la RMSE est souvent la métrique privilégiée en assurance pour les modèles de coût.

Coefficient de Détermination (R^2)

$$R^2 = 1 - \frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2}$$

Avec Y_i valeur observée, $\hat{Y}_i$ valeur estimée et $\bar{Y}$ valeur moyenne

- **Interprétation** : La proportion de la variance de Y expliquée par le modèle. Utile pour l'inférence, mais à utiliser avec prudence pour la prédiction, car il augmente mécaniquement avec le nombre de variables.

Le défi de la classification : l'accuracy et ses limites

- La plupart des problèmes actuariels intéressants (fraude, sinistre grave, résiliation) sont **rares**.
- Ce **déséquilibre des classes** rend la métrique la plus intuitive, **l'accuracy (taux de succès),** totalement trompeuse.
- **Exemple** : Un modèle qui prédit systématiquement "Ne rachète pas" aura une **accuracy de 97%** tout en étant parfaitement inutile pour l'objectif métier (identifier les clients à risque).
- **Conclusion** : Nous devons utiliser des métriques qui se concentrent sur la capacité du modèle à identifier correctement la classe minoritaire (la classe d'intérêt).

Evaluation empirique des scores (1/2)

Matrice de confusion et AUC

Exemple illustratif

Occurrence sinistre	*Score*	*Prédiction (seuil 0.5)*
0	0.4	0
1	0.7	1
1	0.35	0
0	0.52	1
1	0.68	1
0	0.2	0
0	0.1	0
0	0.6	1
1	0.7	1
0	0.55	1

Approche 1 : Matrice de confusion

Mesure la performance des modèles de classification à 2 classes ou plus. Dans le cas binaire, la matrice de confusion est un tableau à 4 valeurs représentant différentes combinaisons de valeurs réelles et valeurs prédites

		Réalité	
		Négative : 0	Positive : 1
Prédiction	Négative : 0	Vrai Négatif	Faux Négatif
	Positive : 1	Faux Positif	Vrai Positif

Evaluation empirique des scores (1/2)

Matrice de confusion et AUC

Exemple illustratif

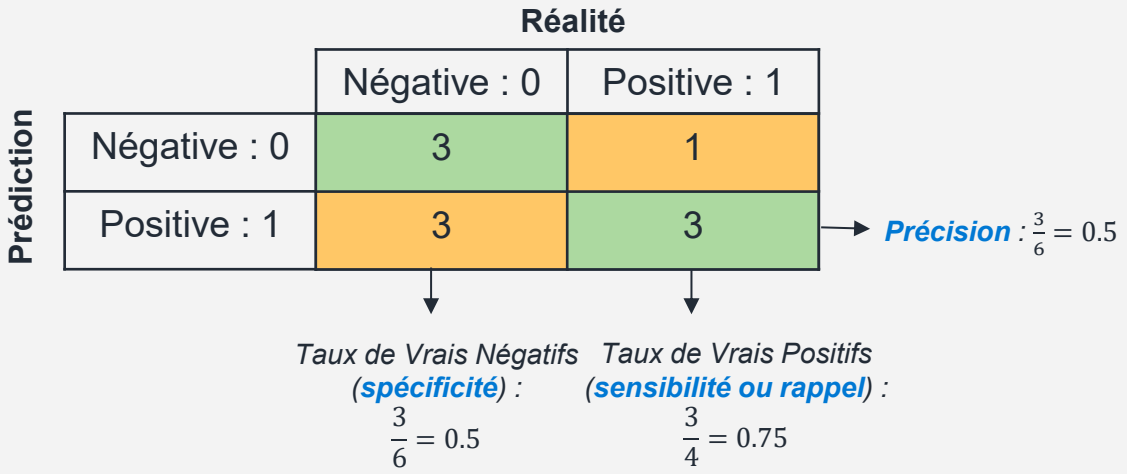
Occurrence sinistre	*Score*	*Prédiction (seuil **0.5**)*
0	0.4	0
1	0.7	1
1	0.35	0
0	0.52	1
1	0.68	1
0	0.2	0
0	0.1	0
0	0.6	1
1	0.7	1
0	0.55	1

Accuracy : 0.6

(TFP, TVP) (0.25, 0.5)

Approche 1 : Matrice de confusion

Mesure la performance des modèles de classification à 2 classes ou plus. Dans le cas binaire, la matrice de confusion est un tableau à 4 valeurs représentant différentes combinaisons de valeurs réelles et valeurs prédites



Evaluation empirique des scores (1/2)

Matrice de confusion et AUC

Exemple illustratif

Occurrence sinistre	*Score*	*Prédiction (seuil 0.5)*	*Prédiction (seuil 1)*
0	0.4	0	0
1	0.7	1	0
1	0.35	0	0
0	0.52	1	0
1	0.68	1	0
0	0.2	0	0
0	0.1	0	0
0	0.6	1	0
1	0.7	1	0
0	0.55	1	0
		Accuracy : 0.6	***Accuracy*** : 0.6
	(TFP, TVP)	(0.25, 0.5)	(0, 0)

Approche 1 : Matrice de confusion

Mesure la performance des modèles de classification à 2 classes ou plus. Dans le cas binaire, la matrice de confusion est un tableau à 4 valeurs représentant différentes combinaisons de valeurs réelles et valeurs prédites

Réalité

Prédiction	Négative : 0	Positive : 1	
Négative : 0	6	4	
Positive : 1	0	0	→ ***Précision*** : NaN

*Taux de Vrais Négatifs (**spécificité**) :* 1

*Taux de Vrais Positifs (**sensibilité ou rappel**) :* $\frac{0}{4} = 0$

ISFA Institut de Science Financière et d'Assurances

Evaluation empirique des scores (1/2)

Matrice de confusion et AUC

Exemple illustratif

Occurrence sinistre	*Score*	*Prédiction (seuil 0.5)*	*Prédiction (seuil 1)*	*Prédiction (seuil **0.65**)*
0	0.4	0	0	0
1	0.7	1	0	1
1	0.35	0	0	0
0	0.52	1	0	0
1	0.68	1	0	1
0	0.2	0	0	0
0	0.1	0	0	0
0	0.6	1	0	0
1	0.7	1	0	1
0	0.55	1	0	0
		***Accuracy** : 0.6*	***Accuracy** : 0.6*	***Accuracy** : 0.9*
	(TFP, TVP)	(0.25, 0.5)	(0, 0)	(0, 0.75)

Approche 1 : Matrice de confusion

Mesure la performance des modèles de classification à 2 classes ou plus. Dans le cas binaire, la matrice de confusion est un tableau à 4 valeurs représentant différentes combinaisons de valeurs réelles et valeurs prédites

Prédiction \ Réalité	Négative : 0	Positive : 1
Négative : 0	6	1
Positive : 1	0	3

→ ***Précision** : 1*

*Taux de Vrais Négatifs (**spécificité**) :* 1

*Taux de Vrais Positifs (**sensibilité ou rappel**) :* $\frac{3}{4} = 0.75$

Indice de Youden pour choisir le meilleur seuil de décision

*TFP (**1 - spécificité**)* — ***À minimiser !***

*TVP (**sensibilité ou rappel**)* — ***À maximiser !***

***Indice de Youden** =* $max \mid TVP - TFP \mid$

ISFA Institut de Science Financière et d'Assurances

Evaluation empirique des scores (2/2)

Matrice de confusion et AUC

Exemple illustratif

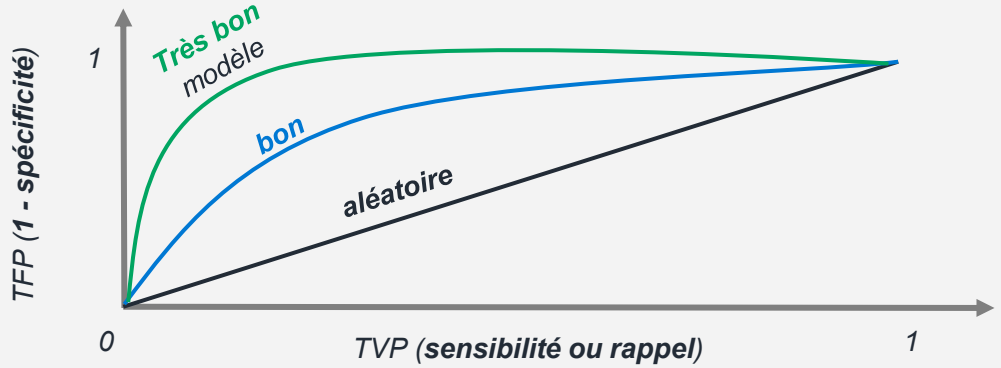
Occurrence sinistre	***Score***	***Prédiction*** *(seuil 0.5)*	***Prédiction*** *(seuil 1)*	***Prédiction*** *(seuil 0.65)*
0	0.4	0	0	0
1	0.7	1	0	1
1	0.35	0	0	0
0	0.52	1	0	0
1	0.68	1	0	1
0	0.2	0	0	0
0	0.1	0	0	0
0	0.6	1	0	0
1	0.7	1	0	1
0	0.55	1	0	0
		Accuracy : 0.6	***Accuracy*** : 0.6	***Accuracy*** : 0.9
	(TFP, TVP)	(0.25, 0.5)	(0, 0)	(0, 0.75)

Approche 1 : Matrice de confusion

Mesure la performance des modèles de classification à 2 classes ou plus. Dans le cas binaire, la matrice de confusion est un tableau à 4 valeurs représentant différentes combinaisons de valeurs réelles et valeurs prédites

Approche 2 : AUC-ROC

L'**AUC-ROC** (Area Under the Curve - Receiver Operating Characteristic) ou **AUC** évalue la capacité d'un modèle à distinguer entre les classes en traçant le TVP contre le TFP **pour différents seuils** de décision. L'AUC (aire sous la courbe) varie entre 0,5 (modèle aléatoire) et 1 (modèle parfait)



Indice de GINI = 2 x AUC - 1

Des mesures **« agnostiques »** aux modèles (s'appliquent à toutes techniques : GLM, arbres de décision, …)

3. Evaluation des modèles et études de cas actuarielles

Les métriques d'évaluation : choisir la bonne règle pour mesurer

- **Études de cas actuarielles :** mettre la théorie en pratique

ISFA
Institut de Science Financière et d'Assurances

Études de cas actuarielles : mettre la théorie en pratique

Nous allons passer en revue trois scénarios : la **tarification**, la **détection de fraude**, et la **prédiction de résiliation**.

1

Tarification en Assurance Automobile

- L'objectif sera de prédire la charge de sinistre attendue, ce qui en fait **un problème de régression**.
- Devons-nous **privilégier un GLM interprétable** pour des raisons réglementaires, ou un modèle plus complexe et performant pour être plus compétitif ?

2

Détection de fraude aux sinistres IARD

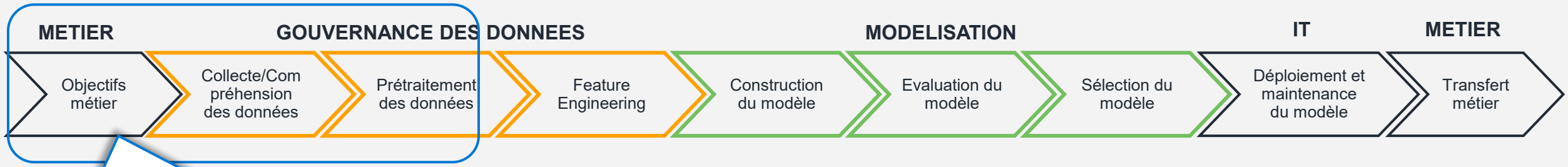
- l'objectif n'est pas de calculer une prime, mais d'optimiser un processus opérationnel en priorisant les dossiers à investiguer.
- Il s'agit d'un problème de **classification binaire**, qui introduit une difficulté technique majeure : le fort **déséquilibre des classes**.

3

Prédiction de la résiliation (*churn*), en assurance vie

- identifier les clients sur le point de partir pour leur proposer des actions de rétention ciblées et rentables.
- C'est également un problème de **classification**, mais avec un **objectif métier différent** de celui de la fraude

Cas 1 - Tarification Auto



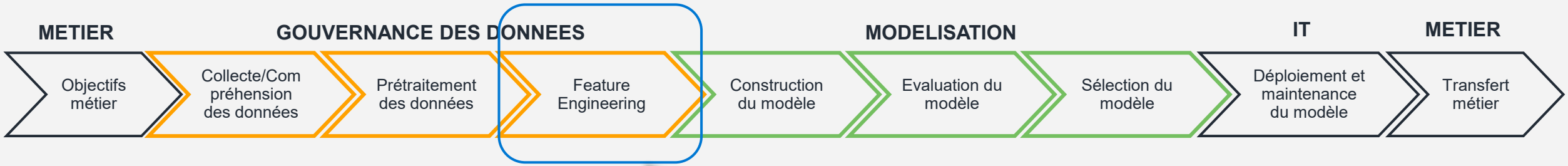
1. Compréhension du besoin métier

- **Objectif** : Améliorer la précision de la tarification pour être plus **compétitif**, tout en maintenant la **rentabilité**. On veut mieux prédire la charge de sinistre annuelle moyenne par contrat.
- **Problème ML** : C'est un problème de régression.

2. Compréhension des données

- Variable Cible (Y) : Charge Totale Annuelle / Exposition.
- Variables Explicatives (X) :
  - *Données Contrat :* Caractéristiques du véhicule (puissance, âge...), garanties souscrites.
  - *Données Assuré :* Âge, genre, ancienneté du permis, profession (CSP), bonus-malus.
  - *Données Géographiques :* Code postal, type de commune (urbain/rural).

Cas 1 - Tarification Auto



1. Compréhension du besoin métier

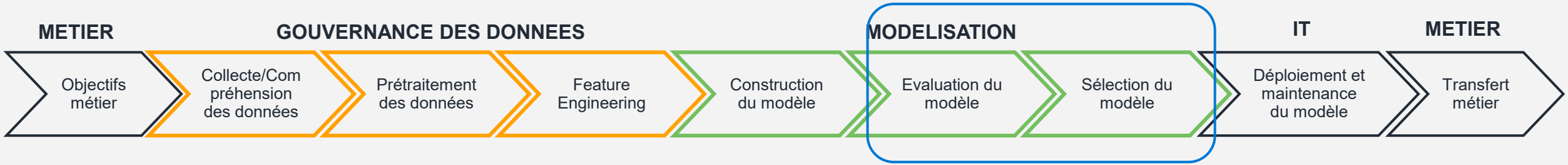
- **Objectif** : Améliorer la précision de la tarification pour être plus **compétitif**, tout en maintenant la **rentabilité**. On veut mieux prédire la charge de sinistre annuelle moyenne par contrat.
- **Problème ML** : C'est un problème de régression.

2. Compréhension des données

- Variable Cible (Y) : Charge Totale Annuelle / Exposition.
- Variables Explicatives (X) :
  - *Données Contrat :* Caractéristiques du véhicule (puissance, âge...), garanties souscrites.
  - *Données Assuré :* Âge, genre, ancienneté du permis, profession (CSP), bonus-malus.
  - *Données Géographiques :* Code postal, type de commune (urbain/rural).

- Création de classes d'âge pertinentes,
- transformation du code postal en variables (densité de population, taux de vol...),
- calcul de l'interaction âge_conducteur * puissance_véhicule.

Cas 1 - Tarification Auto



1. Compréhension du besoin métier

- **Objectif** : Améliorer la précision de la tarification pour être plus **compétitif**, tout en maintenant la **rentabilité**. On veut mieux prédire la charge de sinistre annuelle moyenne par contrat.
- **Problème ML** : C'est un problème de régression.

2. Compréhension des données

- Variable Cible (Y) : Charge Totale Annuelle / Exposition.
- Variables Explicatives (X) :
  - *Données Contrat :* Caractéristiques du véhicule (puissance, âge...), garanties souscrites.
  - *Données Assuré :* Âge, genre, ancienneté du permis, profession (CSP), bonus-malus.
  - *Données Géographiques :* Code postal, type de commune (urbain/rural).

- Création de classes d'âge pertinentes,
- transformation du code postal en variables (densité de population, taux de vol...),
- calcul de l'interaction âge_conducteur * puissance_véhicule.

- **Métriques techniques** : RMSE (pour pénaliser les grosses erreurs de prédiction de sinistres graves) ou MAE.
- **Méthode d'évaluation** : Validation croisée à K plis pour comparer robustement la performance prédictive

Études de cas actuarielles : mettre la théorie en pratique

Nous allons passer en revue trois scénarios : la **tarification**, la **détection de fraude**, et la **prédiction de résiliation**.

1

Tarification en Assurance Automobile

- L'objectif sera de prédire la charge de sinistre attendue, ce qui en fait **un problème de régression**.
- Devons-nous **privilégier un GLM interprétable** pour des raisons réglementaires, ou un modèle plus complexe et performant pour être plus compétitif ?

2

Détection de fraude aux sinistres IARD

- l'objectif n'est pas de calculer une prime, mais d'optimiser un processus opérationnel en priorisant les dossiers à investiguer.
- Il s'agit d'un problème de **classification binaire**, qui introduit une difficulté technique majeure : le fort **déséquilibre des classes**.

3

Prédiction de la résiliation (*churn*), en assurance vie

- identifier les clients sur le point de partir pour leur proposer des actions de rétention ciblées et rentables.
- C'est également un problème de **classification**, mais avec un **objectif métier différent** de celui de la fraude

Laissé en exercice

Conclusion première session

Objectif et Évaluation

Ce que nous avons appris dans cette session :

- **Le "Pourquoi" :** La révolution des données a rendu l'apprentissage statistique indispensable en actuariat, complétant l'inférence par la **prédiction**.
- **La Théorie :** La performance d'un modèle est régie par le **compromis biais-variance**. Le sur-apprentissage est le risque principal.
- **La Méthodologie :** Un processus rigoureux (**CRISP-DM**) et une évaluation robuste (**Validation Croisée**) sont nécessaires pour construire des modèles fiables.
- **La Pratique :** Le choix des **métriques** et l'analyse des résultats doivent toujours être guidés par l'**objectif métier**.

Milliman

ISFA
Institut de Science Financière et d'Assurances
GRANDE ÉCOLE D'ACTUARIAT & GESTION DES RISQUES

Merci

François HU

Francois.hu@milliman.com