ANDRIANARISON Fitiavana
FANOMEZANTSOA Gesy Louis Xavier

M1 Intelligence artificielle et data science

Modélisation Probabiliste en Assurance Automobile

I. **Introduction et Objectifs**

1. **Pourquoi modéliser en assurance automobile ?**

L'assurance automobile doit **« prédire l'imprévisible »** : combien de sinistres vont survenir et combien ils vont coûter. La modélisation probabiliste permet de :

- **Calculer des primes justes :** pour chaque assuré
- **Estimer les réserves** nécessaires pour payer les sinistres futurs
- **Gérer les risques** de l'entreprise d'assurance

2. **Les 3 questions fondamentales**
   - **Le modèle de fréquence** : À quelle fréquence les sinistres surviennent-ils ?
   - **Le modèle de sévérité** : Combien coûte chaque sinistre ?
   - **Le modèle composé** : Quel sera le coût total par assuré ?

ETAPE 1 : Identification des Variables Aléatoires

a. **Variable N : Fréquence des sinistres**

C'est le nombre de sinistres qu'un assuré va déclarer dans l'année

- **Type** : Variable discrète (0, 1, 2, 3, 4, ...)
- **Valeurs possibles** : $N \in \{0, 1, 2, 3, ...\}$
- **Interprétation :** "Combien de fois cet assuré va-t-il avoir un accident cette année ?"

Exemple concret :

- Assuré A : N = 0 (aucun sinistre)

- Assuré B : N = 1 (un sinistre)

- Assuré C : N = 3 (trois sinistres)

b. **Variable X : Coût individuel des sinistres**

C'est le montant d'un sinistre donné

- **Type** Variable continue positive
- **Valeurs possibles** : $X > 0$ (en euros)
- **Interprétation** : "Si un sinistre survient, combien va-t-il coûter ?"

Exemple concret :

- Sinistre léger : X = 800€ (pare-choc rayé)

- Sinistre moyen : X = 5 000€ (collision avec dégâts)

- Sinistre grave : X = 25 000€ (accident avec blessés)

c. **Variable Y : Coût total annuel**

C'est le coût total de tous les sinistres d'un assuré dans l'année

$Y = X_1 + X_2 + ... + X_N$

Interprétation :

- Si N = 0 : Y = 0 (aucun sinistre)

- Si N = 1 : $Y = X_1$ (un sinistre de coût X_1)

- Si N = 2 : $Y = X_1 + X_2$ (deux sinistres)

- Et ainsi de suite...

ETAPE 2 : Modélisation de la Fréquence (Loi de Poisson)

a) **Pourquoi la loi de Poisson ?**

La loi de Poisson est parfaite pour modéliser **des événements rares** qui surviennent **de manière aléatoire** dans le temps :

- Les sinistres automobiles sont relativement rares
- Ils surviennent de manière imprévisible
- Leur survenance est généralement indépendante

b) **Définition mathématique**

Si N ~ Poisson(λ), alors : $P(N = k) = (\lambda^k \times e^{(-\lambda)}) / k!$

Où :

- **k** = nombre de sinistres (0, 1, 2, 3, ...)
- **λ** = paramètre d'intensité (nombre moyen de sinistres par an)
- **e** = constante mathématique (≈ 2.718)
- **k!** = factorielle de k

c) **Interprétation du paramètre λ**

λ représente le nombre moyen de sinistres par assuré et par an

Exemples :

- λ = 0.1 → En moyenne, 1 sinistre tous les 10 ans
- λ = 0.2 → En moyenne, 1 sinistre tous les 5 ans
- λ = 0.5 → En moyenne, 1 sinistre tous les 2 ans

Propriétés

$E[N] = \lambda$ (Espérance = paramètre)

$Var[N] = \lambda$ (Variance = paramètre)

Plus λ est grand, plus la fréquence ET la variabilité des sinistres sont élevées.

Estimation du paramètre λ

Méthode du Maximum de Vraisemblance :

$\hat{\lambda} = (n_1 + n_2 + ... + n_m) / m$

- **$n_1, n_2, ..., n_m$** = nombre de sinistres de chaque assuré
- **m** = nombre total d'assurés

Exemple de calcul :

- 1000 assurés
- 150 sinistres au total
- $\hat{\lambda} = 150/1000 = 0.15$

Interprétation :

En moyenne, chaque assuré a 0.15 sinistre par an.

Calcul des probabilités

Pour λ = 0.2 :

$P(N = 0) = (0.2^0 \times e^{(-0.2)}) / 0! = 1 \times 0.819 / 1 = 81.9\%$

$P(N = 1) = (0.2^1 \times e^{(-0.2)}) / 1! = 0.2 \times 0.819 / 1 = 16.4\%$

$P(N = 2) = (0.2^2 \times e^{(-0.2)}) / 2! = 0.04 \times 0.819 / 2 = 1.6\%$

Interprétation :

- 81.9% des assurés n'auront aucun sinistre
- 16.4% des assurés auront exactement 1 sinistre
- 1.6% des assurés auront exactement 2 sinistres

ETAPE 3 : Modélisation du Coût (Loi Log-Normale)

Pourquoi la loi Log-Normale ?

Les coûts de sinistres ont des caractéristiques particulières :

Toujours positifs (on ne peut pas avoir un coût négatif)

Asymétriques : beaucoup de petits sinistres, peu de gros sinistres

Queue lourde : quelques sinistres très coûteux

Définition mathématique

Si $X \sim \text{LogNormale}(\mu, \sigma^2)$, alors $\log(X) \sim \text{Normale}(\mu, \sigma^2)$

Fonction de densité :

$f(x) = (1 / (x \times \sigma \times \sqrt{}(2\pi))) \times \exp(-((\ln(x) - \mu)^2 / (2\sigma^2)))$

Paramètres :

- **μ** = moyenne du logarithme des coûts
- **σ²**= variance du logarithme des coûts

Interprétation des paramètres

Paramètre μ :

- Plus μ est grand, plus les coûts moyens sont élevés
- μ décale la distribution vers la droite

Paramètre σ² :

- Plus σ^2 est grand, plus la variabilité des coûts est élevée
- σ^2 contrôle la dispersion de la distribution

Moments de la loi Log-Normale

Espérance (coût moyen) :

$E[X] = \exp(\mu + \sigma^2/2)$

Variance (variabilité des coûts) :

$Var[X] = \exp(2\mu + \sigma^2) \times (\exp(\sigma^2) - 1)$

Pourquoi ces formules complexes ?

Propriété clé : Si log(X) suit une loi normale, alors X suit une loi log-normale.

Transformation logarithmique :

- **log(X)** est symétrique → plus facile à analyser
- **X** reste positif → respecte la nature des coûts

Estimation des paramètres

Méthode du Maximum de Vraisemblance :

$\hat{\mu} = (1/n) \times \Sigma \ln(x_i)$

$\hat{\sigma}^2 = (1/n) \times \Sigma (\ln(x_i) - \hat{\mu})^2$

Où :

- $\mathbf{x_i}$ = coût du i-ème sinistre
- **n** = nombre de sinistres observés
- **ln** = logarithme naturel

Exemple de calcul :

Sinistres observés : 1000€, 2500€, 800€, 15000€, 3200€

$\ln(1000) = 6.908$

$\ln(2500) = 7.824$

$\ln(800) = 6.685$

$\ln(15000) = 9.616$

$\ln(3200) = 8.071$

$\hat{\mu} = (6.908 + 7.824 + 6.685 + 9.616 + 8.071) / 5 = 7.821$

Calcul de $\hat{\sigma}^2$:

$\hat{\sigma}^2 = [(6.908-7.821)^2 + (7.824-7.821)^2 + ... + (8.071-7.821)^2] / 5$

$\hat{\sigma}^2 = [0.833 + 0.000 + 1.290 + 3.223 + 0.063] / 5 = 1.082$

Calcul du coût moyen

Avec $\hat{\mu} = 7.821$ et $\hat{\sigma}^2 = 1.082$:

$E[X] = \exp(7.821 + 1.082/2) = \exp(8.362) = 4\ 347€$

Interprétation : Le coût moyen d'un sinistre est d'environ 4 347€.

ÉTAPE 4 : Modèle Composé (Coût Total)

Concept du modèle composé

Le coût total Y combine :

1. **La fréquence** des sinistres (combien ?)

2. **La sévérité** des sinistres (combien chacun coûte ?)

$Y = X_1 + X_2 + ... + X\_N$

Où :

- **N** ~ Poisson(λ) : nombre de sinistres
- **$X_1, X_2, \ldots$** ~ Log Normale (μ, σ^2) : coût de chaque sinistre

Moments du modèle composé

Espérance (coût total moyen) :

$E[Y] = E[N] \times E[X] = \lambda \times E[X]$

Variance (variabilité du coût total) :

$Var[Y] = E[N] \times Var[X] + Var[N] \times (E[X])^2$

$Var[Y] = \lambda \times Var[X] + \lambda \times (E[X])^2$

Interprétation des formules

Espérance $E[Y] = \lambda \times E[X]$:

- Si $\lambda = 0.2$ et $E[X] = 4000€$
- Alors $E[Y] = 0.2 \times 4000€ = 800€$
- **Signification** : En moyenne, chaque assuré coûte 800€ par an

Variance $Var[Y] = \lambda \times Var[X] + \lambda \times (E[X])^2$:

- **Premier terme $\lambda \times Var[X]$** : Variabilité due aux coûts variables
- **Second terme $\lambda \times (E[X])^2$** : Variabilité due à la fréquence variable

Exemple complet de calcul

Données :

- $\lambda = 0.15$ (fréquence)
- $\mu = 8.0$, $\sigma^2 = 1.2$ (paramètres log-normaux)

Étape 1 : Calcul des moments de X

$E[X] = \exp(8.0 + 1.2/2) = \exp(8.6) = 5\,431€$

$Var[X] = \exp(2\times8.0 + 1.2) \times (\exp(1.2) - 1)$

$= \exp(17.2) \times (3.32 - 1)$

$= 26\,903\,134 \times 2.32 = 62\,415\,271€^2$

Étape 2 : Calcul des moments de Y

$E[Y] = 0.15 \times 5\,431 = 815€$

$Var[Y] = 0.15 \times 62\ 415\ 271 + 0.15 \times (5\ 431)^2$

$= 9\ 362\ 291 + 4\ 417\ 854$

$= 13\ 780\ 145€^2$

Écart-type de Y :

$\sigma[Y] = \sqrt{13\ 780\ 145} = 3\ 712€$

Interprétation :

- Coût moyen par assuré : 815€
- Écart-type : 3 712€
- La plupart des assurés coûteront entre 0€ et 4 527€ (815 + 3712)

ÉTAPE 5 : Provisionnement et Réserves

Objectif du provisionnement

L'assureur doit **mettre de l'argent de côté** pour payer :

1. Les sinistres déjà déclarés mais pas encore payés

2. Les sinistres qui vont être déclarés plus tard

3. Les ajustements sur les estimations actuelles

Types de reserves

RBNS (Reported But Not Settled)

Définition : Réserves pour les sinistres déclarés mais non réglés

Calcul simple :

RBNS = Σ (Estimation du coût final - Montant déjà payé)

Exemple :

- Sinistre A : Estimation 5000€, Payé 2000€ → RBNS = 3000€
- Sinistre B : Estimation 3000€, Payé 0€ → RBNS = 3000€
- **Total RBNS = 6000€**

IBNR (Incurred But Not Reported)

Définition : Réserves pour les sinistres survenus mais non déclarés

Problème : Comment estimer des sinistres qu'on ne connaît pas encore ?

Méthode statistique :

IBNR = E[Y] × Nombre d'assurés - Sinistres déjà connus

Exemple :

- 10 000 assurés
- E[Y] = 800€ par assuré
- Coût théorique total = 8 000 000€
- Sinistres déjà déclarés = 6 500 000€
- IBNR = 8 000 000€ - 6 500 000€ = 1 500 000€

IBNER (Incurred But Not Enough Reserved)

Définition : Ajustement car nos estimations peuvent être **insuffisantes**

Méthode simple

IBNER = Pourcentage × (RBNS + IBNR)

Exemple avec 5% d'ajustement :

IBNER = 0.05 × (6 000 000€ + 1 500 000€) = 375 000€

Méthode Chain Ladder (Avancée)

Principe : Analyser comment les sinistres **se développent dans le temps**

Triangle de développement :

Année	Mois 1	Mois 2	Mois 3	Mois 6	Mois 12
2021	100	120	135	145	150
2022	110	130	140	155	???
2023	105	125	138	???	???
2024	115	135	???	???	???

Facteurs de développement :

f_{1-2} = 120/100 = 1.20

f_{2-3} = 135/120 = 1.125

f_{3-6} = 145/135 = 1.074

f_{6-12} = 150/145 = 1.034

Projection pour 2022 :

Mois 12 = 155 × 1.034 = 160€

ÉTAPE 6 : Calcul des Primes

Structure d'une prime d'assurance

Prime commerciale = Prime pure + Chargements

Prime pure

Définition : Coût technique pur du risque

Formule :

Prime pure = $E[Y] = \lambda \times E[X]$

Exemple :

- $\lambda = 0.15$, $E[X] = 4000$€
- **Prime pure = 0.15 × 4000€ = 600€**

Chargements

Les chargements couvrent :

1. **Frais de gestion** (15-20%)

2. **Commissions** (10-15%)

3. **Frais généraux** (5-10%)

4. **Marge de sécurité** (5-10%)

5. **Profit** (5-10%)

Exemple de calcul :

Prime pure : 600€

Chargements : 40% × 600€ = 240€

Prime commerciale : 600€ + 240€ = 840€

Tarification par segment

Principe : Calculer λ_i spécifique à chaque profil d'assuré

Régression de Poisson :

$\log(\lambda_i) = \beta_0 + \beta_1 \times \text{Âge} + \beta_2 \times \text{Sexe} + \beta_3 \times \text{Expérience} + \ldots$

Exemple :

- Jeune conducteur (22 ans, homme) : $\lambda = 0.35$
- Conducteur expérimenté (45 ans, femme) : $\lambda = 0.08$
- **Primes différentes selon le profil**

ÉTAPE 7 : Système Bonus-Malus

Objectif

Ajuster la prime en fonction de l'expérience de sinistralité de l'assuré

Fonctionnement classique

Coefficients standards :

- Coefficient initial : 100%
- Bonus annuel (0 sinistre) : -5%
- Malus (1 sinistre) : +25%
- Malus (2 sinistres) : +50%

Exemple d'évolution :

Année 1 : 100% (début)

Année 2 : 95% (0 sinistre → bonus)

Année 3 : 90% (0 sinistre → bonus)

Année 4 : 112.5% (1 sinistre → malus sur 90%)

Année 5 : 107% (0 sinistre → bonus sur 112.5%)

Calcul de la prime finale

Formule complète :

Prime finale = Prime de base × Coefficient bonus-malus

Exemple :

- Prime de base : 800€
- Coefficient : 90%
- **Prime finale : 800€ × 0.90 = 720€****

ÉTAPE 8 : Validation et Tests

Tests statistiques essentiels

Test du Chi-2 pour la fréquence

Objectif :

Vérifier si les données suivent bien une loi de Poisson

Hypothèses :

- H_0 : Les données suivent une loi de Poisson
- H_1 : Les données ne suivent pas une loi de Poisson

Statistique de test :

$\chi^2 = \Sigma$ ((Observé - Attendu) ² / Attendu)

Test de Kolmogorov-Smirnov pour les coûts

Objectif : Vérifier si les coûts suivent bien une loi log-normale

Principe : Comparer la distribution empirique avec la distribution théorique

Backtesting

Principe : Tester le modèle sur des données historiques

Méthode :

1. Estimer les paramètres sur les années N-3 à N-1

2. Prédire l'année N

3. Comparer prédictions vs réalité

4. Calculer les erreurs de prédiction

Synthèse : Vue d'ensemble du processus

Cycle complet de modélisation

1. Collecte des données

↓

2. Analyse exploratoire

↓

3. Modélisation fréquence (Poisson)

↓

4. Modélisation sévérité (Log-normale)

↓

5. Modèle composé ($Y = X_1 + ... + X\_N$)

↓

6. Calcul des primes

↓

7. Provisionnement

↓

8. Validation et ajustements

Objectifs atteints

Pour l'assureur :

- Primes équitables et compétitives
- Réserves adéquates
- Maîtrise du risque

Pour l'assuré :

- Prime adaptée à son profil de risque
- Couverture garantie en cas de sinistre

Points clés à retenir

1. **La modélisation probabiliste** est la base de toute tarification d'assurance

2. **Chaque formule** a un sens concret et une utilité pratique

3. **La validation** est essentielle pour s'assurer de la qualité du modèle

4. **L'ajustement continu** permet d'améliorer les prédictions

Pour aller plus loin

Sujets avancés

- Modèles de régression avec covariables
- Méthodes de réassurance
- Modèles de dépendance entre fréquence et sévérité
- Intelligence artificielle en actuariat

Applications pratiques

- Tarification automobile
- Assurance habitation
- Assurance santé
- Analyse de rentabilité