Contexte

Crédit Agricole Assurances est une filiale du Groupe Crédit Agricole dédiée à l'assurance, faisant de celui-ci un acteur multi-expert de la bancassurance et le 1er bancassureur en Europe.
Crédit Agricole Assurances regroupe plusieurs entités, dont Predica et Pacifica, qui proposent une large gamme d'assurances aux particuliers, aux exploitants agricoles, aux professionnels et aux entreprises.
Crédit Agricole Assurances s'engage à offrir des solutions innovantes et adaptées aux besoins des clients, tout en favorisant le développement durable et la responsabilité sociale.
Au sein de l'Académie Data Science du groupe, l'objectif est de participer activement à la montée en compétences des collaborateurs, de partager des connaissances et d'identifier de nouveaux usages.

But

Le contrat Multirisque Agricole, géré par Pacifica, est souscrit par les agriculteurs pour sécuriser leur exploitation. Il couvre l'activité professionnelle, les dommages aux bâtiments d'exploitation, le matériel stocké, ainsi que la protection financière et juridique. Ce contrat garantit à l'assuré une couverture efficace et durable, assurant ainsi la continuité de son activité en cas de sinistre, tant sur le plan matériel que financier.

Actuellement, le risque d'incendie constitue une part majeure de la charge sinistre du contrat Multirisque Agricole, ce qui en fait un enjeu clé à modéliser avec précision.

L'objectif est d'identifier le meilleur modèle pour prédire la prime pure incendie, en utilisant :

1. Un modèle pour la Fréquence, (fréquence des accidents pour un profil risque)
2. Un modèle pour le Coût moyen.(CM pour un profil risque)

Prime pure : socle tech du calcul de la prime = estimation stats du cout moyen des sinistres que l'assureur devra en charge. Calcul qui prennent en compte la fréquence, le cout moyen des sinistres pour un profil de risque donné. En gros = CM des sinistre attendus pour un rsk donné

La variable cible finale, la charge, est obtenue en multipliant la fréquence, le coût moyen, et le nombre d'années depuis la souscription du contrat (la variable "ANNEE_ASSURANCE").

Description des données

Un fichier supplémentaire est mis à disposition regroupant toutes les variables disponibles, accompagnées de leur description. Ce fichier inclut :

Les variables cibles : `FREQ`, `CM`, et `CHARGE`

- **Données géographiques** : département, données météorologiques, etc.
- **Données spécifiques au contrat**, notamment :
- L'activité de l'assuré (cultivateur, polyculteur, etc.)

- Les indicateurs de souscription des garanties
- Le nombre de bâtiments, de salariés, et de sinistres déclarés lors de la souscription
  Données de surface : surfaces des bâtiments (élevage, exploitation, etc.), anonymisées en `surface1`, `surface2`, etc., pour garantir la confidentialité
- **Données de capitaux** : capitaux assurés pour différentes options (vol, serres, etc.), anonymisés en `capital1`, `capital2`, etc.
- **Données liées à la prévention** : présence d'équipements (extincteurs, structure en bois, etc.), anonymisées en `prev1`, `prev2`, etc.

Description du benchmark

Objectif du challenge

L'objectif de ce challenge est de comparer les performances des modèles développés dans le cadre de cette compétition avec celles d'un modèle de référence basé sur des **GLM (Generalized Linear Models)** classiques.

Structure du benchmark

Le benchmark repose sur deux modèles GLM distincts :

- **Fréquence des sinistres** :
  - Distribution : *Loi de Poisson*
  - Fonction de lien : *Log*
- **Coût moyen d'un sinistre** :
  - Distribution : *Tweedie*
  - Fonction de lien : *Log*

Évaluation

L'évaluation des modèles repose sur une métrique unique : **RMSE**

L'idée reste d'évaluer dans quelle mesure les approches proposées permettent de dépasser les performances des modèles standards tout en prenant en compte :

- La précision des prédictions
- Les aspects **d'interprétabilité** et d'efficacité
- Les **contraintes métier associées**.

La prime (somme que paie le souscripteur d'un contrat d'assurance) est composée de 3 parties :

- le risque : représentant le coût potentiel du sinistre à assurer ;
- les frais : de gestion par exemple, qui permettent à l'assurance de couvrir ses charges (loyers, salaires des employés) ;
- le bénéfice : soit la marge (positive ou négative) que l'assureur décide d'accorder à une certaine population en lien avec ses objectifs commerciaux. Par exemple, s'il souhaite attirer une population jeune, considérée comme à risques, il acceptera une marge négative sur cette population.

Le calcul risque / prime varie en fonction des critères attribués au type d'assurance. Il n'existe pas de formule de calcul définie pour une prime d'assurance. L'assureur regardera principalement les critères suivants :

- **profil du souscripteur** : âge, état de santé, responsable de sinistres antérieurs ;
- **objet assuré** : habitation, auto, prêt immobilier ;
- **localisation** : les tarifs varient selon les régions, ou si vous vivez en milieu urbain ou rural ;
- **risque couvert** : si les garanties sont nombreuses, la prime d'assurance augmentera ;
- **franchise**, souvent oubliée, influe fortement sur le montant de la prime d'assurance. Plus la franchise d'assurance sera élevée, moins la prime sera importante.