Data Science

M2 Actuariat

Projet de recherche

François HU

Responsable du pôle Intelligence Artificielle, Milliman R&D

Enseignant ISFA

2025

Le projet final : de la Recherche à la pratique Actuarielle (1/4)

Appliquer une méthode de recherche à un cas d'usage assurantiel

Le projet vise à développer votre capacité à effectuer une **veille scientifique**, à assimiler une méthodologie quantitative issue de la recherche, et à en **évaluer l'applicabilité et la pertinence dans le domaine actuariel**.

Phase d'étude théorique

Sélection et **analyse d'un article de recherche** présentant un modèle ou une technique d'apprentissage statistique.

Compétence évaluée :
Rigueur dans la compréhension d'un formalisme mathématique

Phase d'application empirique

Implémentation de la méthode étudiée **sur un jeu de données assurantielles pertinent**.

Compétence évaluée :
Autonomie dans la mise en œuvre algorithmique

Phase d'analyse critique et de synthèse

Évaluation rigoureuse des résultats obtenus, **discussion des apports**, des **limites**, et **formulation de perspectives d'extension**.

Compétence évaluée :
Esprit critique dans l'interprétation des résultats

Le projet final : de la Recherche à la pratique Actuarielle (2/4)

Livrables et structure du rapport

L'évaluation du projet repose sur deux livrables distincts et complémentaires :

1. **Dépôt de code source (via GitHub) :**
   - Un fichier README.md documentant la structure du dépôt et les instructions d'exécution.
   - Les scripts ou notebooks contenant l'implémentation, dûment commentés.
   - Les jeux de données utilisés, à l'état brut et après prétraitement.

2. **Rapport du projet (6 pages maximum, format PDF) :**
   - **Section 1 : Synthèse de la méthodologie de référence (≈ 2 pages)**

     Problématique et objectifs de l'article étudié / Description formelle de la méthode proposée / Résultats et conclusions des auteurs.
   - **Section 2 : Protocole expérimental (≈ 1.5 pages)**

     Présentation du cas d'usage actuariel et du jeu de données / Justification de la pertinence de la méthode pour ce cas d'usage / Description succincte de la méthodologie de prétraitement et de modélisation appliquée
   - **Section 3 : Analyse des résultats (≈ 1.5 pages)**

     Présentation quantitative et qualitative des résultats obtenus / Interprétation de vos résultats
   - **Section 4 : Conclusion et perspectives (≈ 1 page)**

     Analyse critique des limites de l'approche dans le contexte appliqué / Identification de pistes d'amélioration ou d'extensions futures / Référencement d'autres articles pour étayer l'analyse

     Une **annexe (appendix)** peut être ajoutée **sans limite de pages**. Elle est destinée à contenir des résultats complémentaires, des graphiques additionnels ou des résultats mathématiques non essentiels à la compréhension du corps du texte. Le rapport doit rester compréhensible sans la lecture de l'annexe.

Le projet final : de la Recherche à la pratique Actuarielle (3/4)

Organisation, échéance et support

Constitution des Groupes

- Le projet doit être réalisé en groupes de 3 à 4 étudiants.
- La formation des groupes est de votre responsabilité.

Contrainte d'attribution :

- Un même article ne peut être choisi que par **deux groupes au maximum**.
- L'attribution se fera sur la base du **premier arrivé, premier servi**.

Contrainte pour les groupes partageant un sujet :

Les deux groupes travaillant sur le même article devront impérativement utiliser des **datasets/usecases différents** pour leur application.

Sélection du Sujet

- Une **liste d'articles de référence est fournie** à titre indicatif.
- La proposition d'un article externe par un groupe est possible, et doit faire l'objet d'une validation préalable par l'enseignant.

Ressources

- **Jeux de Données :** La collection CAS Datasets (https://cas.uqam.ca/pub/web/CASdatasets-manual.pdf) est recommandée.
- L'utilisation d'autres sources de données est autorisée (Kaggle, UCI, données publiques) si le cas d'usage est bien justifié.

Échéance et Soumission :

- **Date limite Soumission : Samedi 10 Janvier 2026, 23h59.**
- **Procédure de soumission :**

Le rapport au format PDF est à envoyer à francois.hu@milliman.com

L'URL du dépôt GitHub doit être explicitement mentionnée sur la page de garde du rapport.

Aucune soumission tardive ne sera évaluée.

Pour valider le choix de votre groupe et de votre sujet, vous devez envoyer un email (**avant le 31 octobre !**).
Destinataire : francois.hu@milliman.com

- **Objet de l'email** : [Projet Data Science Actuariat] Choix de Sujet
- **Contenu de l'email** : Titre de l'article de recherche choisi. La liste complète des membres du groupe (Nom, Prénom).
- **Important** : Un email de confirmation vous sera envoyé pour valider l'attribution de votre sujet. Si un sujet a déjà été attribué deux fois, il vous sera demandé d'en choisir un autre. Il est fortement recommandé de vous organiser rapidement pour avoir le choix le plus large possible.

Le projet final : de la Recherche à la pratique Actuarielle (4/4)

Catégorie 1 : Méthodes Ensemblistes et traitement des données

- **CatBoost: unbiased boosting with categorical features.** (Prokhorenkova, L., et al., 2018)

Thème : Amélioration du Gradient Boosting pour les **variables catégorielles**.

Contrainte : Comparer le traitement natif des catégorielles de CatBoost à d'autres méthodes (OHE, Target Encoding) sur un jeu de données riche en variables nominales.

- **SMOTE: Synthetic Minority Over-sampling Technique.** (Chawla, N. V., et al., 2002)

Thème : Gestion du **déséquilibre de classes**.

- **mice: Multivariate Imputation by Chained Equations in R.** (Buuren, S. van, & Groothuis-Oudshoorn, K., 2011)

Thème : Imputation des **valeurs manquantes**.

Contrainte : Comparer l'impact de différentes stratégies d'imputation (moyenne, MICE, imputation par k-NN) sur la performance et la robustesse d'un modèle.

- **Practical Lessons from Predicting Clicks on Ads at Facebook.** (He, X., et al., 2014)

Thème : Modèles hybrides et **feature engineering automatique**.

Contribution Clé : Popularise l'architecture GBDT + Régression Logistique, où les arbres servent à créer des features non-linéaires pour un modèle linéaire.

Contrainte : Implémenter ce modèle hybride et le comparer à un GBDT seul et à une Régression Logistique seule sur un problème de classification.

Le projet final : de la Recherche à la pratique Actuarielle (4/4)

Catégorie 2 : Stabilité et sélection de variables

- **Stability selection.** (Meinshausen, N., & Bühlmann, P., 2010)

Thème : Sélection de variables robuste.

- **A new approach to variable importance in random forests.** (Strobl, C., et al., 2008)

Thème : Importance des variables non biaisée pour les modèles en arbres.

Contrainte : Comparer l'importance de variable standard (Gini) à l'importance par permutation, et montrer comment la première peut être trompeuse.

- **The Dantzig selector: Statistical estimation when p is much larger than n.** (Candes, E., & Tao, T., 2007)

Thème : Alternative au Lasso pour la **régression parcimonieuse**.

Contrainte : Implémenter (ou trouver une implémentation) du Dantzig selector et le comparer au Lasso en termes de variables sélectionnées et de performance prédictive sur un jeu de données p > n.

Le projet final : de la Recherche à la pratique Actuarielle (4/4)

Catégorie 3 : Interprétabilité des modèles (XAI)

- **Anchors: High-Precision Model-Agnostic Explanations.** (Ribeiro, M. T., et al., 2018)

Thème : Explications locales par **règles**.

- **Visualizing the effects of predictor variables in black box supervised learning models.** (Apley, D. W., & Zhu, J., 2020)

Thème : Visualisation des **effets de variables non-linéaires** (robuste à la corrélation).

Contrainte : Comparer les graphiques PDP et ALE pour une variable corrélée et analyser les différences d'interprétation.

- **Counterfactual Explanations without Opening the Black Box: A Vector Quantized Autoencoder Approach.** (Poyiadzi, R., et al., 2020)

Thème : Explications **contrefactuelles**.

Contrainte : Implémenter un algorithme de recherche d'explication contrefactuelle (même une version simple) pour un modèle de score de risque.

Le projet final : de la Recherche à la pratique Actuarielle (4/4)

Catégorie 4 : Apprentissage Non Supervisé

- **Isolation-based Anomaly Detection.** (Liu, F. T., et al., 2012)

Thème : Détection d'anomalies.

- **HDBSCAN: Hierarchical Density-Based Clustering.** (McInnes, L., et al., 2017)

Thème : Clustering par densité hiérarchique.

Contrainte : Comparer la robustesse et la qualité des clusters trouvés par K-Means, DBSCAN et HDBSCAN sur un jeu de données complexe.

Le projet final : de la Recherche à la pratique Actuarielle (4/4)

Catégorie 5 : sujets avancés et transversaux

- **Equality of Opportunity in Supervised Learning.** (Hardt, M., et al., 2016)

Thème : Équité algorithmique (Fairness).

- **Entity Embeddings of Categorical Variables.** (Guo, C., & Berkhahn, F., 2016)

Thème : Deep Learning pour Données Tabulaires.

- **TabNet: Attentive Interpretable Tabular Learning.** (Arik, S. Ö., & Pfister, T., 2019)

Thème : Deep Learning pour Données Tabulaires et **Interprétabilité**.

- **Predicting good probabilities with supervised learning.** (Niculescu-Mizil, A., & Caruana, R., 2005)

Thème : Calibration des probabilités.

- **k-Shape: Efficient and Accurate Clustering of Time Series.** (Paparrizos, J., & Gravano, L., 2015).

Thème : Séries temporelles et clustering

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

Merci

François HU

Francois.hu@milliman.com