Introduction

Avec l'essor du big data et des avancées en intelligence artificielle, le secteur de l'assurance connaît une transformation profonde dans ses approches d'analyse et de modélisation des risques. Le machine learning, en particulier, joue un rôle clé dans l'extraction de tendances et la prédiction des sinistres. En adoptant ces méthodes innovantes, les assureurs peuvent affiner leurs modèles prédictifs et améliorer la segmentation des profils d'assurés, ce qui les aide à optimiser la tarification et la gestion des risques.

Objectifs du Machine Learning dans cette Étude:

- Prédiction du Montant des Sinistres: Identifier les modèles les plus précis pour estimer le montant des réclamations d'assurance, en tenant compte de toutes les interactions complexes entre les variables.
- Classification des Profils Risqués: Détecter les profils à risque parmi les assurés pour mieux cibler les efforts de prévention et adapter les politiques de tarification.

Approches Explorées:

1. Régression Ridge:

- Forme régularisée de la régression linéaire, la régression Ridge est utile pour réduire les effets du surajustement.
- Elle s'avère particulièrement efficace lorsque les variables indépendantes présentent une multicolinéarité.

2. Random Forest:

- Ce modèle d'ensemble combine plusieurs arbres de décision pour capturer les interactions complexes et proposer des prédictions robustes.
- Il est résistant au surajustement en raison de la moyenne des prédictions individuelles.

3. XGBoost:

- Un algorithme de boosting performant qui construit progressivement ses modèles pour améliorer la précision.
- Reconnu pour sa capacité à gérer de grands ensembles de données et ses fonctionnalités d'optimisation avancée.

Justification de l'Approche Machine Learning:

1. Complexité des Relations:

- Les données d'assurance présentent des relations non linéaires et multivariées, ce qui limite l'efficacité des modèles linéaires classiques.
- Le machine learning est capable de saisir ces relations complexes grâce à sa flexibilité.

2. Grande Variabilité:

- Les sinistres suivent souvent une distribution très asymétrique, rendant les modèles traditionnels peu adaptés.
- Les modèles de machine learning offrent une meilleure adaptation à ces différentes échelles.

3. Précision et Optimisation:
- Le machine learning propose des méthodes efficaces d'optimisation, telles que la validation croisée et la recherche d'hyperparamètres.
- Cela permet d'affiner les modèles pour obtenir une précision prédictive accrue.

Les techniques de machine learning offrent une opportunité unique d'améliorer la précision des modèles prédictifs et la segmentation des assurés. En combinant différentes approches, il devient possible de répondre aux défis des données complexes tout en obtenant une meilleure gestion des risques.

Le secteur de l'assurance automobile repose sur des modèles prédictifs pour évaluer les risques et fixer les tarifs des polices. L'usage du machine learning (ML) révolutionne cette industrie en offrant une puissance analytique sans précédent, permettant d'extraire des tendances et de prédire les sinistres de manière plus précise. Le présent mémoire explore différentes techniques de ML pour analyser les facteurs qui influencent les montants des sinistres et identifier les assurés les plus susceptibles de présenter des réclamations importantes.

Les données proviennent de la fusion des ensembles contract_data, df_telematics, et claims_data, offrant une vue complète des caractéristiques des contrats, des comportements de conduite et des réclamations d'assurance.
Le dataset a 3791 lignes et 20 colonnes.

Dans cette section, deux études distinctes en machine learning ont été menées pour répondre à des objectifs spécifiques liés à l'analyse des sinistres d'assurance. La première étude vise à prédire les montants des sinistres à l'aide de techniques de régression, tandis que la seconde se concentre sur la classification des assurés selon leur risque de réclamation. Ces analyses permettent de comparer l'efficacité des différents modèles et de mieux comprendre les facteurs clés qui influencent les résultats.

Étude 1: Prédiction du Montant des Sinistres

Objectif: Développer des modèles capables de prédire le montant des sinistres (AMT_Claim) en utilisant des techniques avancées de machine learning.

Approche:
- Préparation des Données:
  - Les variables catégorielles ont été encodées en utilisant OneHotEncoder pour transformer les valeurs en indicateurs binaires.
  - Les données ont été normalisées à l'aide de StandardScaler pour garantir une échelle uniforme entre les prédicteurs.
  - Les données ont ensuite été divisées en ensembles d'entraînement (70 %) et de test (30 %) pour évaluer la performance des modèles.

Modèles Développés:
- Régression Ridge:
  - Une forme de régression linéaire régularisée, Ridge minimise les coefficients des variables en introduisant une pénalité alpha pour réduire le surajustement.
  - GridSearchCV a été utilisé pour optimiser alpha et déterminer la meilleure configuration.
  - Le modèle final a un RMSE de 1923,81, avec des coefficients significatifs identifiant les variables Years.noclaims, Credit_score_cat_Low, et Duration comme principaux prédicteurs du montant des sinistres.

- Random Forest:
  - Un ensemble d'arbres de décision, ce modèle combine des prédictions multiples pour améliorer la robustesse et réduire les erreurs.
  - GridSearchCV a optimisé les paramètres n_estimators et max_depth pour obtenir les meilleures performances.
  - Le modèle a fourni un RMSE de 1849,27, surpassant la régression Ridge.

- XGBoost:
  - Un modèle de boosting performant qui améliore progressivement ses prédictions en pondérant les erreurs précédentes.
  - GridSearchCV a aidé à optimiser le nombre d'estimations (n_estimators), le taux d'apprentissage (learning_rate) et la profondeur (max_depth).
  - Le modèle final a un RMSE de 1878,72, légèrement inférieur au Random Forest.

Interprétation des Résultats:
- Comparaison des Modèles:
  - Le modèle Random Forest offre les meilleures performances, tandis que la régression Ridge reste utile pour interpréter les relations.
  - XGBoost présente une performance intermédiaire mais peut être amélioré avec des données supplémentaires.

Facteurs Clés:
- Years.noclaims et Credit_score_cat_Low restent les prédicteurs les plus significatifs, indiquant que l'historique sans réclamation et le faible score de crédit augmentent le montant des sinistres.

Étude 2: Classification des Assurés

Objectif: Classer les assurés selon leur risque de réclamation, pour mieux cibler les stratégies de prévention et de tarification.

Approche:
- Préparation des Données:
  - AMT_Claim a été binarisé en deux classes (au-dessus ou en dessous de la médiane).
  - Les variables catégorielles ont été encodées en valeurs binaires, et les prédicteurs ont été normalisés.
  - Les données ont été divisées en ensembles d'entraînement (70 %) et de test (30 %).

Modèle Développé:
- Random Forest Classifier:
  - Un modèle basé sur la combinaison d'arbres de décision, optimisé via GridSearchCV pour n_estimators et max_depth.
  - Le modèle a été évalué à l'aide d'une matrice de confusion, d'un rapport de classification et d'une courbe ROC.
  - Les prédictions sur l'ensemble de test ont donné une précision de 0,61 et un AUC de 0,67, indiquant une performance modérée.

Interprétation des Résultats:

- Matrice de Confusion:
  - La précision montre que le modèle est capable d'identifier correctement environ 61 % des assurés dans chaque classe.
  - Cependant, des améliorations sont nécessaires pour réduire les erreurs de classification.

- Courbe ROC:
  - La courbe ROC révèle un compromis entre les taux de vrais et faux positifs, avec un AUC de 0,67, indiquant une capacité de distinction modérée.

Les deux études montrent le potentiel du machine learning pour prédire et classer les sinistres. Bien que les modèles de régression Ridge et Random Forest offrent des résultats intéressants, des améliorations sont nécessaires pour réduire les erreurs et améliorer la classification. L'intégration de données supplémentaires et l'optimisation des modèles sont des pistes à explorer pour de meilleurs résultats futurs.

L'analyse économétrique, réalisée en première étape, a permis d'identifier des tendances générales et d'isoler les principaux facteurs affectant le montant total des sinistres (AMT_Claim). Cette approche fournit des modèles interprétables et permet de mettre en lumière les relations directes entre les variables, telles que Duration et Years.noclaims.

Cependant, ces modèles présentent des limites lorsqu'il s'agit de saisir des relations non linéaires complexes. C'est là que le machine learning intervient, en offrant une plus grande flexibilité et en capturant des interactions subtiles entre les variables. En utilisant des techniques telles que le Random Forest et XGBoost, il est possible d'améliorer les prédictions et d'obtenir des modèles plus performants, tout en conservant la capacité de dégager des insights précieux pour la gestion des risques.

Limites et Recommandations

Limites Identifiées:
1. Qualité des Données:
  - Certaines variables ne sont pas toujours disponibles ou sont mal documentées, limitant leur intégration dans les modèles.
  - Des valeurs aberrantes subsistent dans certaines variables comme AMT_Claim.

- La fusion des dataframes a considérablement réduit le nombre d'observation. Il faudrait donc trouver un moyen de faire nos prédictions en gardant intact le nombre d'observations d'origine.

2. Performances Prédictives:
  - Les modèles présentent des erreurs relativement élevées dans leurs prédictions, ce qui laisse penser que d'autres facteurs restent non identifiés.
  - L'algorithme XGBoost, malgré son efficacité, n'atteint pas les niveaux de précision attendus dans la prédiction.

3. Multicolinéarité:
  - Les modèles économétriques souffrent de multicolinéarité, faussant les coefficients et la significativité des variables.

Recommandations:

1. Amélioration des Modèles:
  - Optimiser davantage les hyperparamètres des modèles par validation croisée pour améliorer leur précision.
  - Explorer d'autres modèles non linéaires tels que le Neural Network pour des prédictions complexes.

2. Enrichissement des Données:
  - Intégrer plus de données données externes (par exemple, estimer AMT_Claim pour les autres observations qui n'en avaient pas) pour améliorer les prédictions.

3. Segmentation et Tarification:
  - Utiliser le machine learning pour segmenter les profils d'assurés en groupes homogènes, afin d'adapter les politiques tarifaires.
  - Mettre en place des politiques de prévention ciblant les segments les plus risqués.

Conclusion Générale

L'analyse combinée de l'économétrie et du machine learning offre une vision globale et détaillée des facteurs influençant les sinistres d'assurance. Les techniques économétriques permettent d'identifier les relations linéaires et d'extraire les principales variables, tandis que le machine learning complète cette analyse par la découverte d'interactions complexes.

Les modèles prédictifs développés démontrent le potentiel des nouvelles technologies pour améliorer la gestion des risques dans le secteur de l'assurance. Néanmoins, il reste des défis à relever, notamment en matière de qualité des données et de multicolinéarité. Des recommandations claires sont proposées pour dépasser ces obstacles et adapter les stratégies tarifaires aux profils d'assurés identifiés, permettant ainsi une tarification plus précise et une prévention ciblée.