UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

GUIFT DIMITRI **2024-2025**
12214381

MASTER 2 MoSEF DATA SCIENCE

NOTE DE SYNTHÈSE

CREDIT AGRICOLE ASSURANCES

ASSURANCES

Interprétabilité des clauses contractuelles en assurances dommages et modélisation du sinistre tempête

Tuteur d'alternance : M. Mustapha BENARBIA

Unité d'accueil - lieu d'alternance
Ville : Paris Pays : France

Tuteur pédagogique : M. Marc-Arthur DIAYE

Note de synthèse

Depuis le 30 septembre 2024, j'ai l'opportunité de réaliser mon alternance au sein de Crédit Agricole Assurances, au sein de la Direction de l'Audit des Assurances (DAA) située à Paris Montparnasse. Cette direction constitue la troisième ligne de défense du Groupe et assure un rôle stratégique dans la gouvernance de l'entreprise. Elle conduit des missions visant à évaluer le degré de maîtrise des risques techniques, financiers, opérationnels, informatiques ou encore de conformité. Dans ce contexte, le pôle Data, auquel j'ai été rattaché, a été créé afin d'accompagner la transformation digitale de l'audit. Sa mission est d'exploiter les données à grande échelle, de développer des outils analytiques et de mettre en place des méthodes de data science pour renforcer la pertinence et l'efficacité des travaux. Mon rôle de data scientist s'inscrivait donc directement dans cette dynamique, avec pour objectif de mettre en œuvre des approches innovantes relatives principalement au machine learning au service de l'audit interne.

Au cours de cette expérience, plusieurs missions m'ont été confiées, chacune répondant à des enjeux métiers concrets. Le premier projet concernait l'interprétabilité des clauses contractuelles en assurance dommages. Les contrats d'assurance contiennent parfois des formulations ambiguës qui peuvent susciter des litiges entre l'assureur et l'assuré. L'objectif était de concevoir un modèle automatique capable d'identifier ces clauses interprétables afin de sécuriser la rédaction des contrats, de renforcer la transparence pour les clients et d'éviter d'éventuelles sanctions de régulateur comme l'ACPR (Autorité de contrôle prudentiel et de résolution). Concrètement, ce travail a impliqué la constitution d'un jeu de données de clauses, leur prétraitement linguistique et la mise en œuvre de modèles supervisés tels que la régression logistique, le Random Forest ou encore le Gradient Boosting. Les résultats obtenus ont montré la faisabilité d'une telle approche, mais ont également mis en lumière des défis importants liés à la complexité du langage juridique et à la taille limitée du corpus. D'un point de vue métier, ces travaux ouvrent la voie à un outil d'aide à la rédaction qui permettrait de réduire les zones d'ambiguïté et de limiter les risques de contentieux. Pour aller plus loin, l'utilisation de modèles avancés de Hugging Face (choses qui n'étaient pas disponibles en entreprise), adaptés au traitement du langage en français, représenterait une amélioration pertinente.

Le second projet portait sur le risque climatique appliqué à l'assurance agricole. Dans un contexte où les événements extrêmes, tels que les tempêtes, deviennent plus fréquents et plus intenses, la capacité à anticiper la sinistralité constitue un enjeu crucial pour les assureurs. Mon travail a consisté à construire une base de données robuste en croisant le portefeuille Multirisque Agricole

avec les données climatiques issues d'ERA5. Cette base, structurée à la maille contrat–trimestre, a servi de support à la modélisation de la probabilité d'un sinistre lié au vent. Plusieurs algorithmes ont été testés, parmi lesquels la régression logistique, le Random Forest, LightGBM et CatBoost. Les résultats ont permis d'identifier des variables clés comme l'intensité des rafales ou l'historique de sinistres, confirmant la pertinence de l'approche. Toutefois, l'analyse a été confrontée à un déséquilibre marqué entre les contrats sinistrés et non sinistrés, nécessitant l'utilisation de techniques de rééchantillonnage. D'un point de vue métier, cette démarche illustre la possibilité de doter les assureurs d'outils prédictifs pour mieux calibrer leurs garanties et anticiper la sinistralité climatique lié au vent. Des prolongements pourraient inclure l'intégration d'autres aléas (sécheresse, inondation, gel) et l'utilisation de méthodes hybrides combinant expertise actuarielle et apprentissage automatique.

Ces deux projets mettent en évidence la valeur ajoutée de la data science pour l'audit interne. Ils montrent qu'il est possible de passer de contrôles ponctuels et manuels à des analyses massives et automatisées, capables de cibler rapidement les zones à risque. Néanmoins, ils soulignent aussi les conditions nécessaires à leur réussite : la qualité des données, la taille des bases et la collaboration étroite entre experts métier et data scientists. Sur le plan personnel, cette alternance m'a permis de consolider mes compétences techniques en modélisation statistique, en machine learning et en traitement automatique du langage, tout en développant une compréhension fine des problématiques métier de l'assurance dommage surtout. J'ai également appris à travailler en étroite collaboration avec des auditeurs, ce qui m'a permis de progresser dans ma capacité à vulgariser mes analyses et à les inscrire dans une logique opérationnelle.

En définitive, cette alternance représente une étape déterminante de mon parcours. Elle m'a permis d'évoluer dans un environnement exigeant, à l'interface entre audit et sciences des données, et de contribuer à des projets concrets porteurs d'impact pour l'entreprise. Elle m'a montré que l'apport de la data science ne se limite pas à la performance des modèles, mais réside dans sa capacité à répondre à des besoins réels et à renforcer la gouvernance.