Mélanie Yakoub

Note de Synthèse

Au cours de mon alternance au sein du Groupe Crédit Agricole, j'ai participé activement à deux projets majeurs en traitement automatique du langage naturel (NLP). Ces projets ont non seulement enrichi mes compétences techniques en data science et en intelligence artificielle, mais ont également contribué à l'optimisation des processus de gestion des risques opérationnels au sein de la Direction des Risques du Groupe.

Projet 1 : Cartographie des Risques Opérationnels et Informatiques

Le premier projet sur lequel j'ai travaillé consistait à développer une cartographie des risques opérationnels et informatiques en utilisant des techniques avancées de NLP. Ce projet avait pour objectif de regrouper et d'analyser les descriptions textuelles d'incidents opérationnels afin d'extraire des thèmes récurrents et de détecter de nouvelles zones de risque. Cette cartographie visait à améliorer la gestion des risques au sein du Groupe Crédit Agricole en offrant à l'équipe une meilleure compréhension des différents risques et en permettant ainsi une meilleure définition des catégories de risques et des ajustements stratégiques.

La première étape de ce projet consistait à nettoyer de manière approfondie les données textuelles disponibles. Ce nettoyage était crucial pour garantir la qualité des analyses futures. Les descriptions d'incidents contenaient souvent des erreurs typographiques, des abréviations, et d'autres anomalies qui pouvaient nuire aux résultats. J'ai donc mis en place un pipeline de nettoyage des données qui incluait la suppression des caractères spéciaux, des chiffres et de la ponctuation, la normalisation des termes, et la correction des fautes d'orthographe, l'elimination des mots vides (dit stopword) et le stemming et lemmatisation qui permet de reduit les mots à leur formes caniques ou leur racines et creee une methode de detection des abrviations les plus importantes. Ce processus a permis d'améliorer la qualités des données mais n'a pa été suffisant pour traiter toutes les abreviations.

Ensuite, j'ai appliqué différentes méthodes de vectorisation textuelle pour transformer ces descriptions en vecteurs numériques exploitables. Deux techniques principales ont été utilisées : le TF-IDF (Term Frequency-Inverse Document Frequency) et BERTopic qui exploite le modele BERT (Bidirectional Encoder Representations from Transformers). Le TF-IDF m'a permis d'évaluer l'importance des termes dans chaque document, tandis que BERT, grâce à ses capacités de compréhension contextuelle, a permis une représentation plus riche et nuancée des textes.

J'ai ensuite exploré plusieurs algorithmes de clustering pour regrouper les incidents en clusters thématiques. J'ai notamment utilisé K-Means, le clustering hiérarchique, et HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise). Après avoir comparé les résultats à l'aide de l'indice de silhouette, j'ai opté pour une approche combinée : le TF-IDF avec K-Means pour une vue d'ensemble des clusters, et BERTopic (basé sur BERT et HDBSCAN) pour une analyse plus fine et précise des thèmes émergents. Une fois les clusters formés, j'ai effectué une analyse approfondie en utilisant le CountVectorizer pour identifier les termes les plus fréquents dans chaque cluster. Cela m'a permis de mieux comprendre les thèmes récurrents au sein des incidents, en retraçant les descriptions et en isolant les vecteurs associés à chaque groupe.

Le résultat final de ce projet a été l'intégration de ces méthodes dans une application interactive. Cette application permet aux experts en risques du Groupe Crédit Agricole d'explorer et d'analyser les clusters identifiés, améliorant ainsi la compréhension et la gestion des risques opérationnels, tout en offrant un outil puissant pour la détection des anomalies et l'identification de nouvelles catégories de risques.

Projet 2 : Prédiction des Catégories de Risques

Le deuxième projet, qui a constitué la majeure partie de mon alternance au sein du Groupe Crédit Agricole, portait sur la prédiction des catégories de risques opérationnels (ET3) à partir des descriptions d'incidents. Ce projet avait pour objectif d'automatiser la classification des incidents dans les sous-catégories de risques, une tâche traditionnellement effectuée manuellement par les experts et souvent sujette à des erreurs, en raison de la complexité liée aux 200 catégories existantes. L'enjeu principal était d'améliorer la précision et la fiabilité de cette classification, tout en permettant aux experts de gagner en efficacité.

Pour ce faire, j'ai d'abord adapté les techniques de nettoyage de données que j'avais développées lors du premier projet, en les ajustant aux besoins spécifiques des modèles de prédiction utilisés. En proposant différents niveaux de nettoyage en fonction des exigences de chaque modèle, j'ai pu optimiser la qualité des données, ce qui est crucial pour la performance des algorithmes de machine learning.

Afin de surmonter le déséquilibre des classes présentes dans les données, où certaines catégories de risques étaient surreprésentées, j'ai mis en place des techniques de rééchantillonnage telles que SMOTE et ajusté les métriques de performance. Cela a permis d'améliorer l'équilibre et la robustesse des prédictions. J'ai ensuite exploré plusieurs techniques de vectorisation pour transformer les descriptions textuelles en vecteurs numériques exploitables. En plus des méthodes TF-IDF et BERT, j'ai expérimenté avec Word2Vec pour capturer les relations sémantiques entre les mots, enrichissant ainsi la représentation des données textuelles.

Une fois les descriptions d'incidents vectorisées, j'ai appliqué divers algorithmes de machine learning, tels que les forêts aléatoires, les réseaux de neurones, ainsi que des modèles de fine-tuning basés sur BERT, pour classifier les incidents. J'ai optimisé ces modèles de machine learning à l'aide de techniques d'optimisation comme la recherche par grille (Grid Search) et l'optimisation bayésienne, puis j'ai comparé les différents résultats pour sélectionner les trois meilleurs modèles selon les performances obtenues. Afin d'améliorer encore les performances, ces modèles ont été affinés en introduisant une sous-sélection par une catégorie générale de risque.

Vectorisation	**Modèle ML**	**Nettoyage**	**Sous sélection de variables**	**Caisse et période**	**Accuracy**
TF IDF	Random Forest	Avancé	Non	Ile de France	92%
Bert	Random Forest	Léger	Non	Ile de France/ 2018 à 2022	88%
Bert	Finetuning	Léger	Non	Ile de France	87%
Word2Vec	Random Forest	Avancé	Non	Ile de France	87%
TF IDF	Random Forest	Avancé	Oui	Ile de France	99%
Bert	Random Forest	Léger	Oui	Ile de France	76%
Bert	Finetuning	Léger	Oui	Ile de France	72%

Table 1: Tableau comparatif des performances des modèles ML selon la vectorisation et le nettoyage

Les modèles sélectionnés ont ensuite été intégrés dans une application web que j'ai développée avec Streamlit. Cette application offre une interface utilisateur intuitive, permettant aux experts en gestion des risques de classifier les incidents en proposant diverses options de nettoyage, de sélection de modèles, et d'ajustement des paramètres pour s'adapter aux besoins spécifiques des utilisateurs.

En parallèle, j'ai veillé à ce que l'outil développé soit facilement réutilisable par mes collègues. J'ai créé un package pour regrouper toutes les fonctionnalités développées et dockerisé l'ensemble du projet afin d'en garantir la reproductibilité. Ce projet représente un enjeu significatif pour la gestion des risques au sein de la Direction des Risques du Groupe Crédit Agricole, offrant un outil performant et adaptable pour améliorer la classification des incidents de risques. Cependant, il reste avant tout une expérimentation, car la qualité des données constitue encore une limite importante pour l'intelligence artificielle. Pour atténuer cette limite et améliorer l'utilité de mon application, j'ai inclus une fonctionnalité affichant les cinq classes les plus probables en sortie, plutôt que de ne proposer qu'une seule prédiction, afin de mieux guider les experts.

Compétences Développées

Au cours de ces deux projets, j'ai non seulement approfondi mes connaissances en NLP et en machine learning, mais j'ai également développé des compétences pratiques en gestion de projet et en développement d'applications web. J'ai appris à travailler avec des outils tels que Power BI pour la visualisation des données, Streamlit pour le développement d'applications web, ainsi que Docker et MLflow pour la comparaison des modèles et la containerisation des applications. De plus, j'ai acquis une compréhension approfondie des défis liés au nettoyage des données textuelles et à la gestion des déséquilibres de classes, ainsi que des compétences avancées en modélisation et en optimisation de modèles. Ces projets m'ont permis d'apporter des idées novatrices dans le cadre d'une expérimentation visant à améliorer les processus de gestion des risques opérationnels au sein du Groupe Crédit Agricole, tout en renforçant mes compétences en data science et en traitement du langage naturel. Je suis convaincue que ces expériences seront précieuses pour mes futures missions professionnelles.