OPENCLASSROOMS

CentraleSupélec

Développez une preuve de concept: Détection d'anomalies en assurances

Table des matières

I. Introduction

a. Contexte

Pendant mon parcours de « Machine Learning Engineer » chez Openclassrooms, j'ai réalisé une alternance tout au long de cette année. J'ai effectué ma mission d'entreprise au sein du groupe BNP PARIBAS dans l'entité Cardiff qui couvre le secteur de l'assurance.

Au cours de mon alternance chez BNP Cardiff, j'ai été affecté au département de la Production financière au sein de l'unité commerciale Données Far, faisant partie de l'équipe Data Analyse et Innovation. Notre équipe avait pour mission principale de soutenir les équipes métier, telles que les actuaires, le contrôle de gestion, la comptabilité et l'équipe de risque, dans leur transition vers des processus automatisés et numériques.

Dans l'ensemble de cette mission en entreprise, notre rôle principal était de récupérer et transformer les données. Nous avons mis en place divers outils pour traiter les données provenant d'acteurs externes. Nous avons géré un volume important de données financières et déployé nos propres solutions numériques. De plus, nous avons également formé les équipes métier à l'utilisation des outils que nous avons développés, afin de renforcer leurs compétences.

b. Problématique

Cependant, nous avons constaté que les données que nous recevions présentaient de nombreuses anomalies et incohérences lors du chargement. Le contrôle manuel de ces anomalies était très chronophage et coûteux. Pour remédier à cela, nous avons envisagé une approche automatisée pour effectuer une première vérification de la qualité des données et détecter les éventuelles anomalies.

Une partie importante de ma mission consistait donc à mettre en place un système de contrôle de qualité des données afin de détecter et signaler les lignes de données anormales. Dans ce but, nous avons opté pour une approche basée sur la machine Learning. Nous avons développé des modèles de détection d'anomalies spécifiquement adaptés aux données financières de notre base de données. L'objectif était d'anticiper les anomalies futures que les équipes métier devraient examiner, tout en accélérant et automatisant le processus.

Ainsi, en utilisant ces modèles de détection d'anomalies, nous cherchions à identifier les incohérences et anomalies potentielles dans les données financières, afin de faciliter la tâche des équipes métier et d'améliorer l'efficacité globale du processus.

II. Méthodologie

a. Plan chronologique de l'études

Voici un plan chronologique représentant les grandes étapes de notre travail de détection d'anomalies :

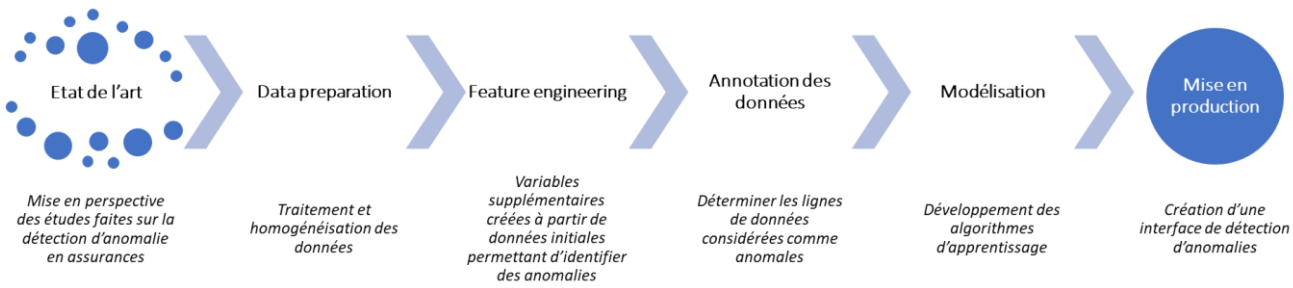

Figure 1- Plan études

1. **État de l'art :** Synthèse des études et travaux sur la détection d'anomalies en assurance et mise en perspective des modèles utilisés.

2. **Préparation des données :** Nous avons regroupé les données provenant de différentes sources et les avons harmonisées. Cette étape comprend la fusion de différentes tables, la gestion des formats de données, la normalisation des valeurs, et toute autre manipulation nécessaire pour préparer les données à être utilisées dans notre étude.

3. **Features Engineering :** Nous avons créé de nouvelles variables (features) à partir des données existantes et sélectionné les variables les plus importantes pour notre modèle. Cela implique l'application de techniques telles que la transformation et la création de variables dérivées pertinentes pour la détection des anomalies.

4. **Annotation des données :** Nous avons organisé des ateliers avec des experts des domaines financiers, comptables et actuariat afin de bénéficier de leur expertise dans l'identification des anomalies dans leurs propres jeux de données. Ces séances de travail ont permis de recueillir des informations précieuses sur les critères de détection des anomalies, en utilisant leur méthode de repérage comme base pour annoter les données existantes.

5. **Modélisation :** Nous avons appliqué des modèles de machine learning à notre situation en utilisant les données préparées et les variables sélectionnées. Cela inclut le choix des algorithmes appropriés, la division des données en ensembles d'entraînement et de test, l'ajustement des hyperparamètres, et l'évaluation des performances des modèles pour la détection des anomalies.

6. **Mise en production :** Une fois que nous avons sélectionné le modèle le plus performant, nous l'avons récupéré et mis en place une interface de déploiement conviviale. Cette interface permet aux utilisateurs d'appliquer le modèle pour détecter les anomalies dans les données en fournissant les entrées nécessaires et en lançant le processus de détection. La mise en production garantit que le modèle est prêt à être utilisé de manière pratique et efficace dans un environnement opérationnel.

b. Etat de l'art

Lors de notre étude, nous avons consulté différents articles afin d'approfondir nos connaissances sur la détection des anomalies. Notre objectif principal était d'identifier les modèles potentiels que nous pourrions exploiter dans ce projet. Voici les modèles qui ont été retenus dans l'état de l'art, en se basant sur les recherches que nous avons effectuées :

1. Mémoire intitulé *"Détection d'anomalies via l'apprentissage non-supervisé : application à la fraude" de Assan Aziz Coulibaly* (Coulibaly, 2021): Ce mémoire explore l'utilisation de méthodes d'apprentissage non-supervisé pour la détection d'anomalies, en mettant l'accent sur l'application spécifique à la fraude. Les méthodes d'apprentissage non-supervisé telles que l'Isolation Forest, le K-means, le C-means et le LOF sont abordées dans ce mémoire.

2. Article intitulé "*Machine Learning techniques for anomaly Detection: An overview*" (Ngadi, 2013): Cette recherche fournit un aperçu des techniques de machine learning utilisées pour la

détection des anomalies. Les méthodes classiques de machine learning telles que les arbres de décision sont discutés dans cet article.

3. Article intitulé *"Anomaly Detection with Machine Learning"* (Johnson, 2020): Cette recherche se concentre spécifiquement sur la détection des anomalies à l'aide de techniques de machine learning. Il aborde différentes approches et méthodes utilisées dans ce domaine.

4. Mémoire intitulé *« Gradient Boosting algorithm for early detection of unknow internet of thing devices »* : Ce mémoire met en avant l'utilisation de l'algorithme de gradient boosting pour la détection précoce des dispositifs inconnus de l'Internet des objets (IoT). Il souligne l'efficacité de cette méthode dans le domaine de l'assurance (Ferman, 2021).

Sur la base de ces recherches, nous avons identifié plusieurs modèles pertinents pour notre projet. Ces modèles incluent l'Isolation Forest, le K-means, le C-means, le LOF, les arbres de décision et l'algorithme de gradient boosting. Ils seront pris en considération lors du développement de notre solution de détection des anomalies. Nous avons implémenté un modèle de classification dummy pour servir de référence dans la comparaison avec d'autres modèles. Ce modèle de classification dummy est relativement simple et se base sur des règles aléatoires ou basiques pour effectuer des prédictions. Il nous permet d'évaluer les performances des autres modèles en les comparant à cette référence. En utilisant le modèle de classification dummy, nous sommes en mesure de déterminer si les autres modèles sont plus performants que ce modèle de base et si leur utilisation est justifiée dans notre contexte.

Modéles upervisés	Modéles non-supervisés
Decision trees	K-means
Random forest	Isolation forest
Gradient Boosting	Self-organizing maps (SOM)
Modèles linéaires	C-means
Multi-layer perceptron	Adaptive resonance theory (ART)
Support vector machine Learning	Local outlier factor (LOF)

Figure 2- Modèles retenus

III. Traitement des données

a. Données

Nous avions accès à une base de données contenant des informations sur les sinistres et les primes. Les primes correspondent aux versements mensuels des assurés et les sinistres correspondent aux montants versés par les assurances quand il y a un litige. Cependant, cette base de données était composée de plusieurs sources non homogènes, provenant de différents délégataires externes. Notre tâche consistait à concaténer et homogénéiser ces données pour les rendre cohérentes et uniformes.

Nous disposons des bases de données internes récupérées qui contiennent des données financières sensibles. Dans le cadre du projet, ces données ont été analysées puis modifiées. Une valeur de coefficient aléatoire unique a été appliquée à chaque donnée.

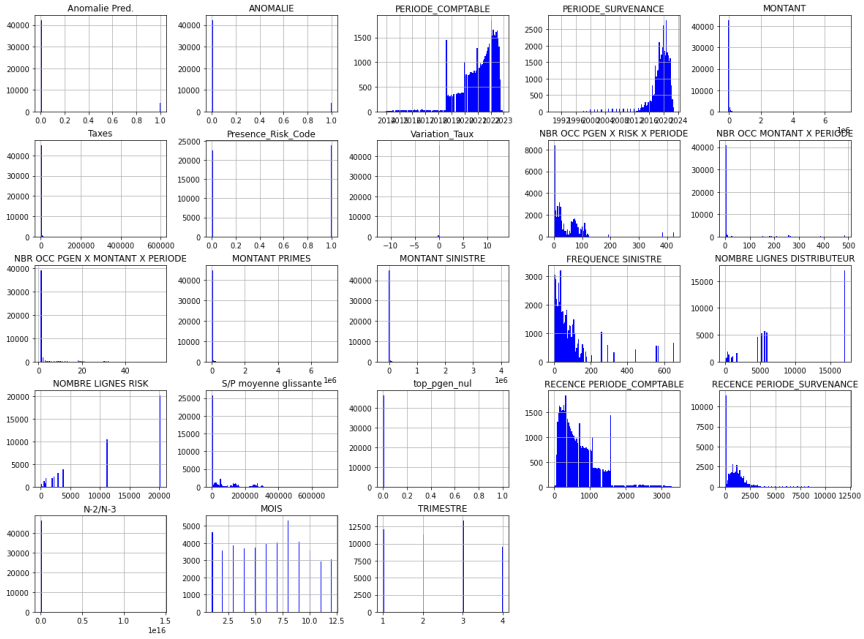

Figure 3- Visualisation des données

b. Feature engineering

Dans le but de faciliter la détection des anomalies dans le jeu de données, nous avons créé des variables supplémentaires. L'ajout de ces variables supplémentaires vise à améliorer les performances des modèles utilisés pour la détection des anomalies.

Features
La variation : N/N-1
La saisonalité : (N-2)-(N-1) /2 (N-3)-(N-1) /2 sur 12 derniers mois par risque (la moyenne de deux mois sur trois mois glissants)
La saisonalité : (N-2)-(N-1) /2 (N-3)-(N-1) /2 sur 12 derniers mois par distributeur (la moyenne de deux mois sur trois mois glissants)
Prime périodique/unique
Indicateur Nouveau PGEN 0/1
Nombre d'occurrence d'un même PGEN x RISK sur une période donnée
Nombre d'occurrence d'un même montant sur une période donnée
Nombre d'occurrence d'un même montant sur une période donnée et un PGEN donné
PGEN vide
Risque présent ou absent du référentiel APLE
Fréquence de sinistres
Fréquence de sinistre remboursés

Figure 4- Features supplémentaires

Ensuite pour faciliter la détection des anomalies dans notre jeu de données, nous avons entrepris différentes étapes pour sélectionner les features les plus pertinentes. Voici les étapes que nous avons suivies :

- ✓ **Data exploration :** Nous avons analysé la distribution des différentes features de notre jeu de données. Cette étape nous a permis d'avoir un aperçu global des caractéristiques présentes et de déterminer celles qui étaient moins pertinentes d'un point de vue technique. Nous avons consulté

les avis des experts seniors de notre équipe pour prendre des décisions éclairées sur les features à conserver.

- ✓ **Analyse en composantes principales (ACP)** : Nous avons appliqué l'ACP pour évaluer la quantité de variance expliquée par nos features. L'objectif était de réduire le nombre de features tout en conservant une quantité significative d'information. Nous avons constaté qu'avec 16 features, nous pouvions expliquer environ 85% de la variance totale de notre jeu de données.

- ✓ **Matrice de corrélation :** Nous avons examiné la corrélation entre les différentes features afin d'éliminer les redondances et de réduire davantage le nombre de features. L'idée était de conserver les features les plus informatives et de minimiser la redondance dans notre modèle.

À la fin de cette étape, nous avons réussi à réduire le nombre de features de 43 à 16.

En suivant ces différentes étapes, nous avons pu sélectionner un ensemble plus restreint mais plus pertinent de features pour notre modèle de détection des anomalies. Cette approche nous permet de maximiser l'efficacité et la performance de notre modèle en se concentrant sur les aspects les plus significatifs du jeu de données.

d. Annotations des Anomalies

Les équipes métiers ont défini les règles d'annotation suivantes :

1. **Annotation des données :** Avec Utilisation de Power BI pour identifier les points considérés comme des anomalies en termes de points. En passant par des méthodes d'identification statistiques.
2. **Règle de gestion :** Mise en évidence des cas où les PGEN (paramètres généraux) sont manquants. Mise en évidence du RISK manquant dans la ligne.

Ces règles ont été établies afin d'identifier et de traiter les anomalies dans les données dans les bases. Voici la visualisation des données annotés comme anomalies.

Voici une visualisation des anomalies :

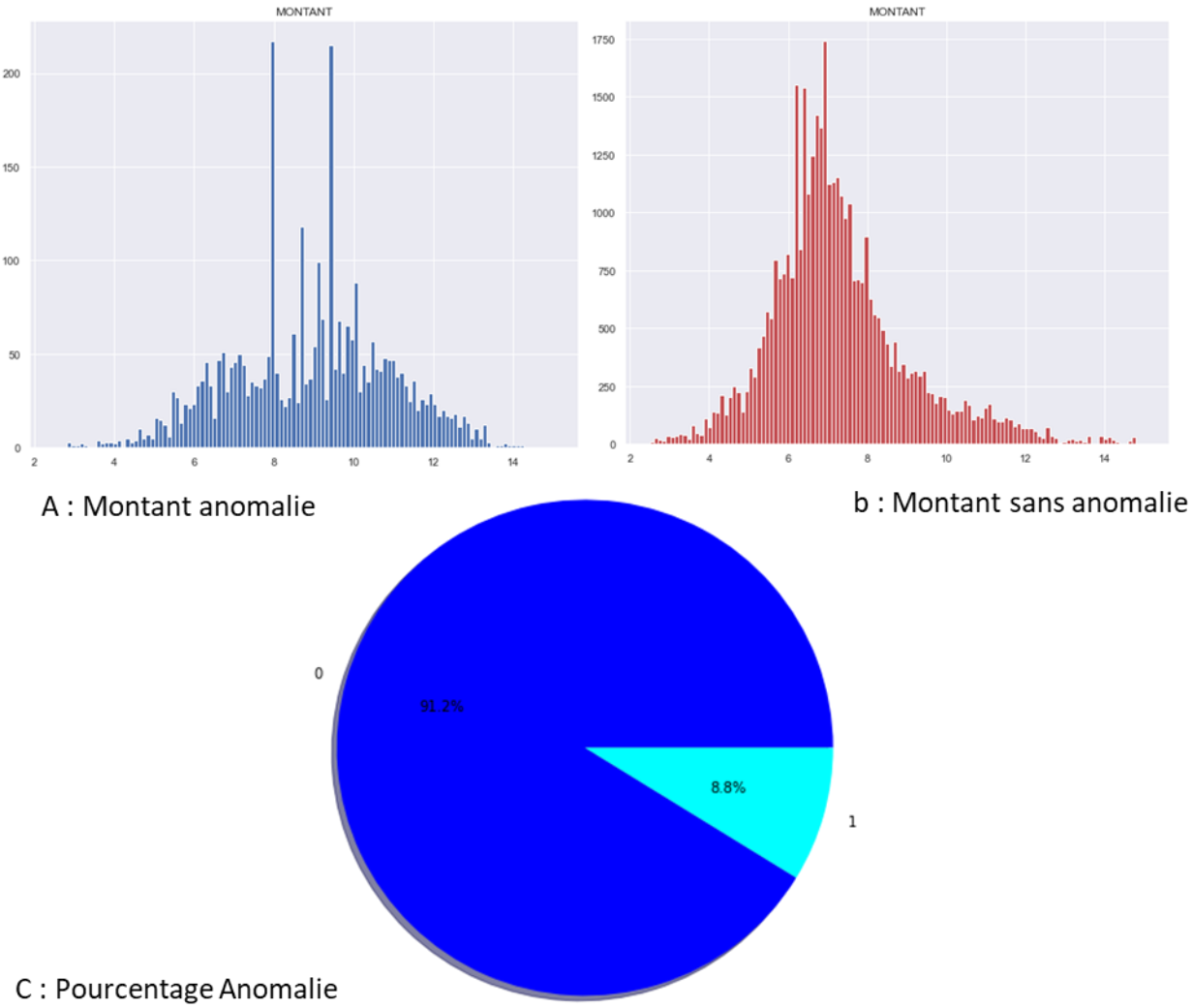

Figure 5 - Visualisation des anomalies

Sur le graphique, nous pouvons constater plusieurs observations. Premièrement, environ 8.8% des données sont des anomalies.

Deuxièmement, nous avons appliqué une transformation logarithmique aux montants des anomalies. Cette transformation permet de mieux mettre en évidence les différences entre les montants des anomalies et ceux des données non anormales.

En examinant les montants, nous pouvons remarquer que les anomalies présentent généralement des montants plus élevés que les données non anormales. Cela suggère que les anomalies se manifestent par des valeurs extrêmes ou inhabituelles dans le jeu de données.

De plus, lorsque nous examinons la distribution des montants des données sans anomalie, nous observons une tendance à suivre une distribution log-normale. Cela signifie que les montants des données normales sont généralement répartis de manière symétrique autour d'une valeur centrale et suivent une croissance exponentielle après avoir été transformés logarithmiquement.

En revanche, les anomalies ne suivent pas cette tendance log-normale. Elles se distinguent par des montants nettement plus élevés et une distribution moins symétrique. Cela renforce l'idée que les anomalies sont des observations inhabituelles et distinctes du reste des données.

Ces observations fournissent des indications précieuses pour la détection des anomalies dans notre jeu de données et suggèrent des stratégies potentielles pour améliorer la performance des modèles de détection.

IV. Choix du modèle et mise en production

a. Modélisation

Nous avons exploré différentes approches, à la fois supervisées et non supervisées, pour détecter les anomalies dans notre jeu de données. Dans un premier temps, nous avons mis en place des modèles de machine learning non supervisés. Malheureusement, les résultats obtenus n'étaient pas concluants et nécessitaient une optimisation plus approfondie des algorithmes. En raison des contraintes de temps, nous avons pris la décision de ne pas poursuivre l'expérience sur les algorithmes non supervisés.

Nous nous sommes alors concentrés sur les approches supervisées, afin de maximiser nos chances de détecter les anomalies. Nous avons commencé par mettre en place des modèles linéaires et des SVM, mais les résultats obtenus n'étaient pas satisfaisants. Nous avons ensuite exploré les approches basées sur les arbres de décision et le gradient boosting, qui se sont avérées nettement plus performantes.

Ensuite, nous avons utilisé un réseau de neurones, qui s'est avéré être une approche prometteuse pour la détection d'anomalies. Cependant, la mise en place et l'optimisation de ce modèle a demandé beaucoup de temps et d'efforts supplémentaires.

Il est important de noter que notre choix final a été basé sur une combinaison de performances et de contraintes de temps. Nous avons opté pour le modèle de détection des anomalies qui présentait les meilleurs résultats parmi ceux que nous avons explorés, tout en prenant en compte les ressources nécessaires pour son optimisation.

En conclusion, notre étude nous a permis de comparer différentes approches de détection des anomalies, tant supervisées que non supervisées. Nous avons choisi de privilégier les modèles supervisés, tels que les arbres de décision et le gradient boosting, en raison de leurs performances supérieures. Toutefois, il convient de noter que l'optimisation des modèles non-supervisés et l'exploration d'autres approches supervisées restent des pistes intéressantes pour de futurs travaux.

b. Résultats

Après avoir testé plusieurs algorithmes de machine Learning dans notre étude de cas, nous avons analysé les résultats obtenus. Parmi les différents modèles évalués, celui qui s'est révélé le plus performant est le modèle de Random forest. Par conséquent, nous avons décidé de le retenir et de le mettre en production pour la détection des nouvelles anomalies.

	fit_time	score_time	test_accuracy	test_f1	test_precision	test_recall	train_accuracy	train_f1	train_precision	train_recall	Model
0	11.766948	0.029199	0.867945	0.553917	0.572834	0.786416	0.996402	0.976283	0.989049	0.963951	XG Boost
0	0.217104	0.032600	0.922808	0.000000	0.000000	0.000000	0.922808	0.000000	0.000000	0.000000	Ridge
0	42.412970	0.092601	0.920481	0.001172	0.202439	0.000622	0.923197	0.010146	0.994872	0.005127	MLP
0	8.633304	0.061000	0.894572	0.707224	0.651495	0.906468	1.000000	1.000000	1.000000	1.000000	Random Forest

Figure 6- Score modèle retenu

Après avoir comparé les différents modèles, nous avons constaté que le Random Forest obtenait les meilleurs résultats. Bien que le modèle de Gradient Boosting aurait pu être une alternative viable, le Random Forest a montré une performance supérieure. Dans notre étude, notre objectif principal était d'optimiser la « précision » afin de détecter le plus grand nombre d'anomalies possible tout en minimisant les faux positifs. Cela nous permet d'avoir une marge de sécurité et de capter potentiellement plus d'anomalies que ce qui est réellement présent dans les bases de données.

	Modèle	Résultat	Validé
Modèle Supervisé	Ridge	Relation trop complexe	X
	Support Vector Machine	Relation trop complexe	X
	Multi-Layer Perceptron	Temps d'execution et d'optimisation trop long	X
	Gradient Boosting	Très bon résultat	X
	Random Forest	Meilleur résultat	O
Modèle non-supervisé	K-Means	Regroupement des petits et grands montants	X
	Isolation Forest	Isolation des sinistres et des primes	X

Figure 7- Tableau de comparaison des modèles.

d. Probabilité des prédictions

Dans notre approche, nous avons cherché à améliorer les performances de notre modèle en ajustant le seuil de probabilité de classification. Par défaut, le seuil de probabilité est fixé à 0,5, ce qui signifie que les échantillons avec une probabilité supérieure à 0,5 pour la classe 1 sont considérés comme des anomalies, et inversement pour la classe 0. En ajustant le seuil de probabilité, nous cherchons à trouver le bon équilibre entre la qualité de détection et la signification des détections de notre modèle.

	0	1
275	0.98977	0.01023
276	0.98977	0.01023
277	0.89977	0.10023
278	0.98977	0.01023
297	0.98977	0.01023
...	...	...
4539	0.69000	0.31000
4540	0.79000	0.21000
4541	0.68000	0.32000
4542	0.68000	0.32000
4543	0.68000	0.32000

539 rows × 2 columns

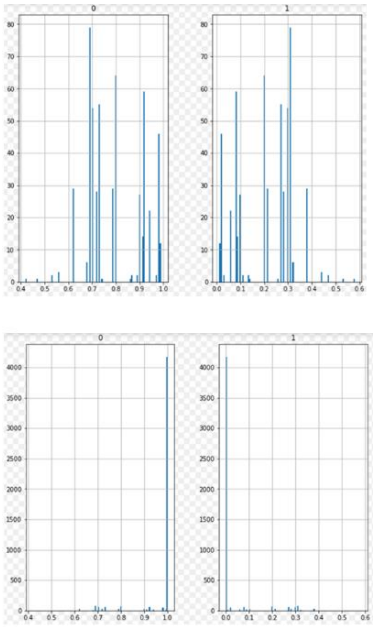

Figure 8 - Probabilité des prédictions

Une observation importante est que nous avons constaté une faible proportion de probabilités comprises entre 0 et 1 dans notre modèle. Cela signifie que le fait de modifier le seuil de probabilité n'a pas eu un impact significatif sur nos résultats.

Cependant, nous avons mené des expérimentations avec différents seuils de probabilité, allant de 0,3 à 0,6. Nous avons observé que les différences entre les seuils de 0,6 et 0,4 étaient minimes, car le nombre de valeurs de classe 1 reste proche. Par conséquent, les scores obtenus restent similaires. En revanche, lorsque le seuil a été fixé à 0,3, le score a considérablement diminué, car le nombre de valeurs de classe 1 a augmenté de manière significative. Globalement, le changement de seuil n'a pas eu un impact majeur sur les scores.

Cela suggère que notre modèle est relativement stable par rapport au seuil de probabilité, et que les performances de détection des anomalies ne sont pas fortement influencées par de petits ajustements du seuil. Les résultats obtenus sont cohérents dans une certaine plage de seuils.

Il est important de noter que ces observations sont spécifiques à notre jeu de données et au modèle utilisé. Dans d'autres cas, il peut y avoir une plus grande variation des performances en fonction du seuil de probabilité. Il est donc recommandé de mener des expérimentations spécifiques pour trouver le seuil optimal en fonction du contexte et des objectifs de détection d'anomalies.

	test_precision	test_recall	Nb_0	Nb_1	Seuil
0	0.662345	0.894236	4142	393	0.6
1	0.651495	0.906468	4136	399	0.5
2	0.648470	0.917468	4122	413	0.4
3	0.716035	0.821318	3771	582	0.3

Figure 9 - Résultat changement de seuil

e. Mise en production

Nous avons ensuite mis notre modèle en production. Nous avons développé une interface permettant de détecter les anomalies dans les bases de données réelles. L'utilisateur peut spécifier la période sur laquelle il souhaite détecter les anomalies, puis il lui suffit de cliquer sur le bouton "Détecter les anomalies" pour lancer le programme. Cette interface simplifiée facilite l'utilisation du modèle et permet aux utilisateurs de bénéficier des fonctionnalités de détection des anomalies facilement utilisable.

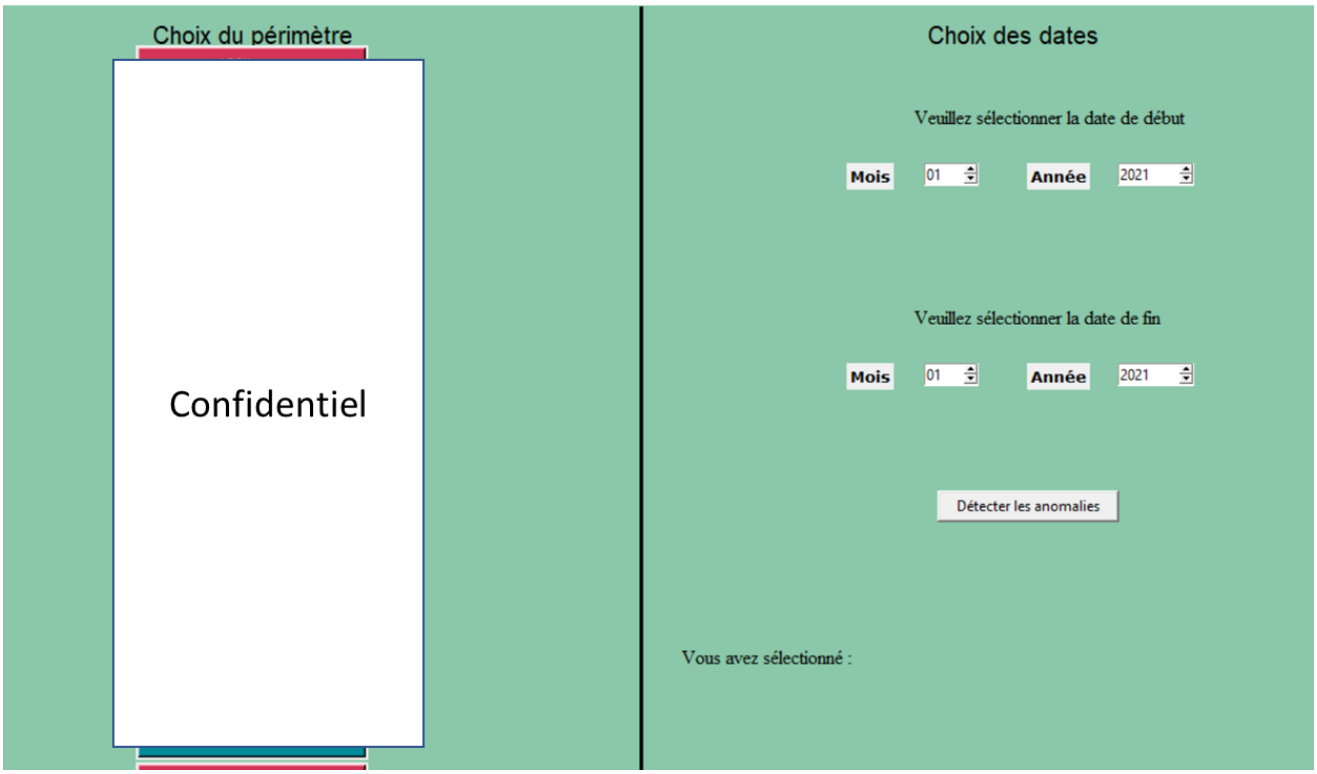

Figure 10 - Capture d'écran de l'interface

Une fois que l'interface a terminé la détection des anomalies, elle génère un fichier Excel contenant toutes les lignes de données. Une colonne supplémentaire intitulée « ANOMALIE PREDITE » est

ajoutée, affichant les valeurs correspondant aux lignes identifiées comme des anomalies. Ce fichier Excel fournit ainsi une vue détaillée des données avec une indication claire des anomalies détectées, ce qui facilite l'analyse ultérieure et les actions correctives nécessaires.

ANOMALIE PREDITE
0
0
1
0
0
0
1
0
0
0
1
0

Figure 11- Capture des anomalies prédites

Voici une représentation graphique des anomalies détectées dans notre jeu de données. En observant les graphiques, on constate que les anomalies ont généralement des montants plus élevés que les autres lignes comptables. De plus, on remarque que les anomalies sont plus fréquentes au deuxième et troisième trimestre de l'année, correspondant aux mois d'avril et de septembre. Cette visualisation met en évidence des tendances intéressantes et peut aider à mieux comprendre la nature des anomalies.

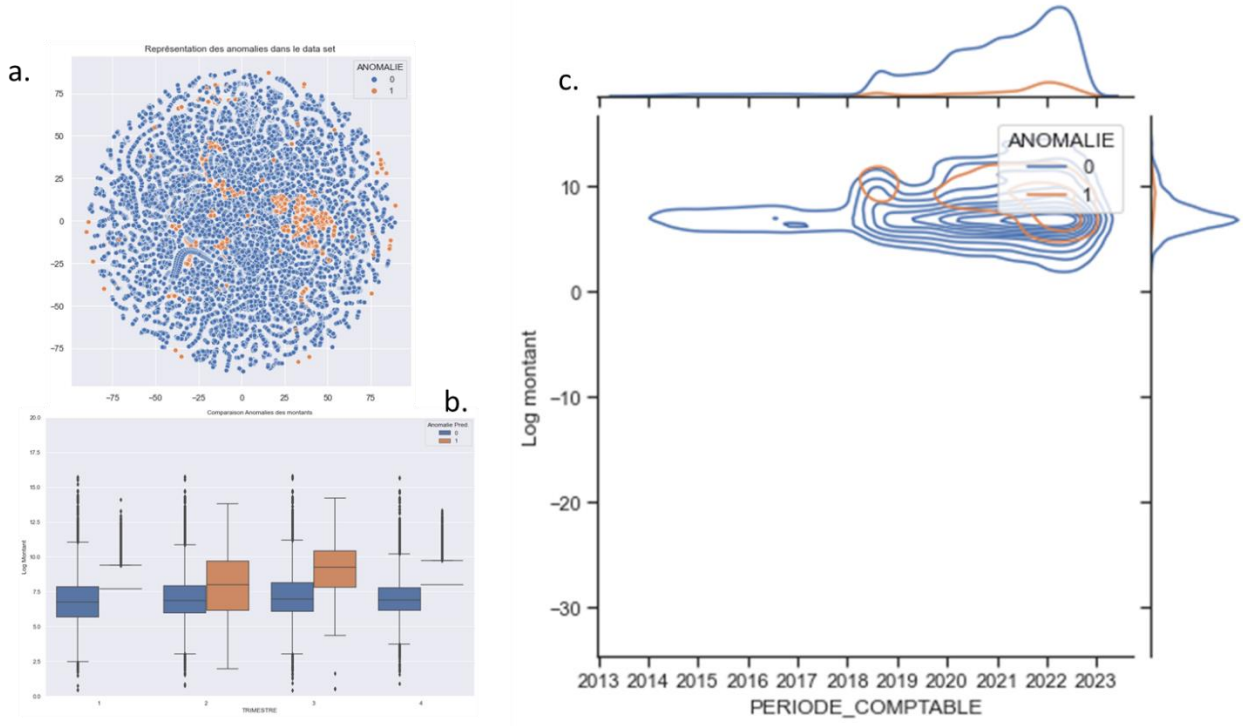

Figure 12- Visualisation des anomalies prédites

V. Synthèse

a. Conclusion

Notre étude avait pour objectif de développer un outil permettant de détecter les anomalies dans un jeu de données afin d'améliorer sa qualité. Dans cette optique, nous avons exploré différentes approches de machine learning. Malheureusement, les approches non supervisées n'ont pas donné les résultats escomptés, ce qui nous a poussés à nous tourner vers les approches supervisées.

Afin de simplifier le traitement des données, nous avons réduit le jeu de données en sélectionnant les caractéristiques les plus pertinentes. Toutefois, nous avons réalisé que la qualité des modèles et le feature engineering avaient un impact significatif sur les résultats obtenus. Par conséquent, nous avons consacré une attention particulière à ces aspects.

Grâce à ces efforts, nous avons pu déployer un modèle de machine learning capable de détecter un maximum d'anomalies, qui peuvent ensuite être évaluées par les experts métier. Cette approche nous a permis de maximiser l'efficacité du processus de détection des anomalies et de faciliter la prise de décision pour améliorer la qualité des données.

En résumé, notre travail a abouti au développement d'un outil basé sur le machine learning, qui offre la possibilité de détecter et évaluer les anomalies dans un jeu de données, contribuant ainsi à l'amélioration globale de sa qualité.

b. Axes d'amélioration.

En conclusion, notre étude sur la détection des anomalies dans les données d'assurance a permis de mettre en évidence plusieurs axes d'amélioration pour optimiser la performance des modèles de détection. L'interface de détection des anomalies doit être améliorée afin de faciliter l'interprétation des résultats et de permettre une prise de décision plus efficace.

Le **feature engineering** est un aspect clé à explorer davantage. En développant de nouvelles variables pertinentes, il est possible d'améliorer la sensibilité et la spécificité de la détection des anomalies. Différentes techniques de feature engineering doivent être explorées pour maximiser les résultats.

L'**approche non supervisée** a été sous-exploitée dans notre étude. En investissant plus de temps et de ressources dans cette approche, nous pourrons découvrir de nouvelles formes d'anomalies et améliorer la détection dans des cas où les schémas d'anomalies ne sont pas bien connus.

Un **apprentissage continu des nouvelles anomalies** est essentiel pour suivre l'évolution des données au fil du temps. En mettant en place un processus d'apprentissage continu, nos modèles pourront s'adapter aux nouvelles anomalies émergentes, assurant ainsi une détection précise et actualisée.

Enfin, l'utilisation de **modèles distincts pour les sinistres et les primes** est recommandée. Ces deux types de données présentent des caractéristiques distinctes, et en utilisant des modèles spécifiques pour chaque type, nous pourrons capturer plus précisément les particularités et les schémas propres à chaque variable, améliorant ainsi la détection des anomalies.

En poursuivant ces axes d'amélioration, nous pourrons maximiser la détection des anomalies dans les données d'assurance, permettant ainsi une gestion plus efficace des risques et une prise de décision éclairée.

En mettant en œuvre ces différentes recommandations, il sera possible d'améliorer la détection des anomalies dans les données d'assurances, de rendre le processus plus robuste, plus efficace et mieux adapté aux spécificités du domaine. Cela contribuera à une meilleure gestion des risques et à une prise de décision plus éclairée pour les acteurs du secteur.

IV. Références

Coulibaly, A. A. (2021). *Détection d'anomalies via l'apprentissage non-supervisé : application à la fraude.* Brest.

Ferman, V. A. (2021). Récupéré sur https://www.researchgate.net/publication/354746373_GRADIENT_BOOSTING_ALGORITHM_FOR_EARLY_DETECTION_OF_UNKNOWN_INTERNET_OF_THINGS_DEVICES

Johnson, J. (2020). Récupéré sur https://www.bmc.com/blogs/machine-learning-anomaly-detection/

Ngadi, S. O. (2013). Récupéré sur https://www.researchgate.net/profile/Salima-Benqdara/publication/325049804_Machine_Learning_Techniques_for_Anomaly_Detection_An_Overview/links/5af3569b4585157136c919d8/Machine-Learning-Techniques-for-Anomaly-Detection-An-Overview.pdf