PREDICTION DE LA FRAUDE DOUANIERE A PARTIR DE L'APPRENTISSAGE MACHINE ET L'ANALYSE MIROIR AU TOGO

PREDICTING CUSTOMS FRAUD USING MACHINE LEARNING AND MIRROR ANALYSIS IN TOGO

PREVISÃO DA FRAUDE ADUANEIRA COM BASE EM APRENDIZADO DE MÁQUINA E ANÁLISE ESPELHO NO TOGO

Pouwemdéou Tchila[*]*; Komlan Kawa Agbanho*[†]*; Abalo Bouwe*[‡]

Résumé

La fraude douanière est un phénomène inhérent aux administrations douanières qui compromet le plus souvent la collecte des recettes douanières. Pour lutter contre ce phénomène, les administrations douanières à fortiori dans les pays en développement effectuent souvent des contrôles intrusifs de façon massive et anarchique. Ce qui n'encourage pas la fluidité du commerce international. L'objectif de cette étude est d'analyser dans quelle mesure le recours à l'apprentissage machine et à l'analyse miroir améliore l'identification des fraudes douanières, tout en préservant l'objectif de mobilisation des recettes. À partir des données issues de l'Office Togolais des Recettes et celles issues de COMTRADE, les résultats montrent que l'analyse miroir et l'apprentissage machine permettent de mieux cibler la fraude douanière. Pour ce faire, l'étude recommande l'utilisation de ces outils dans la détection de la fraude.

Mots clés: machine learning, analyse miroir, fraude douanière

Codes JEL: C55, F14, H26

Abstract

Customs fraud is an inherent phenomenon of customs administrations and is most often responsible for undermining customs revenue collection. In an attempt to combat this phenomenon, customs administrations, particularly in developing countries, often conduct extensive and unstructured audits. This is not conducive to the fluidity of international trade. The objective of this study is to analyse the extent to which the use of machine learning and mirror analysis improves the identification of customs fraud, while preserving the objective of revenue mobilisation. Using data from the

[*] Docteur en sciences économiques, Data scientiste, Chef division analyse risques et suivi-évaluation, Office Togolais des Recettes & Chercheur associé au CREAMO (Université de Lomé), ORCID: 0000-0002-3586-0524, samuelson902@gmail.com

[†] Docteur en sciences économiques, Inspecteur des Douanes, Chef section brigade à la Division des Opérations Douanières de Kwadjoviakopé de l'Office Togolais des Recettes.

[‡] Master en Statistiques & Mathématiques, Data scientiste, Chargé de l'analyse des risques et de la programmation fiscale, Office Togolais des Recettes.

https://doi.org/10.47348/AMTJ/V5/i1a1

Togolese Revenue Authority and COMTRADE, the findings indicate that mirror analysis and machine learning can better enhance customs fraud detection. To this end, the study recommends the use of these tools in fraud detection.

Keywords: machine learning, mirror analysis, customs fraud

Resumo

A fraude aduaneira é um fenómeno inerente às administrações aduaneiras que compromete, na maioria das vezes, a cobrança de receitas aduaneiras. Para combater este fenómeno, as administrações aduaneiras, sobretudo nos países em desenvolvimento, realizam frequentemente controlos intrusivos de forma massiva e anárquica. Isto não favorece a fluidez do comércio internacional. O objetivo deste estudo é analisar em que medida o recurso à aprendizagem automática e à análise espelho melhora a identificação de fraudes aduaneiras, preservando ao mesmo tempo o objetivo de mobilização das receitas. Com base nos dados do Serviço Togolês de Receitas e nos dados da COMTRADE, os resultados mostram que a análise espelho e a aprendizagem automática permitem direcionar melhor a fraude aduaneira. Para tal, o estudo recomenda a utilização destas ferramentas na deteção da fraude.

Palavras-chave: Aprendizagem automática, análise espelho, fraude aduaneira

I. Introduction

Depuis quelques années, les volumes des marchandises franchissant les frontières connaissent de plus en plus une croissance exponentielle du fait de la libéralisation commerciale. En effet, la libéralisation des échanges s'est renforcée ces dernières années grâce à l'Organisation Mondiale du Commerce (OMC). De nos jours, à l'ère de la mondialisation, il est crucial pour les pays et les entreprises de fournir des biens et des services dans les délais et à faible coût. D'une part, les frais de transport élevés, les exigences documentaires complexes, les formalités et les délais de dédouanement aux frontières nationales peuvent faire obstacle à la compétitivité d'un pays et par conséquent, à sa participation dans l'économie mondiale. D'autre part, les contrôles physiques systématiques des marchandises par les administrations douanières peuvent aussi être un frein à la facilitation des échanges. Si certaines études ont souligné que la facilitation est favorable au développement économique (Zaki, 2011; Viet, 2015), il n'en demeure pas moins qu'elle soit entravée par les contrôles physiques anarchiques aux frontières surtout dans les pays en développement. Le principal argument en faveur de ces contrôles est le risque de perte de recettes, cruciales pour le financement du développement économique.

Au Togo, les données du Système douanier automatisé (SYDONIA WORLD) montrent que le nombre de déclarations de marchandises

https://doi.org/10.47348/AMTJ/V5/i1a1

émises est passé de 230 504 déclarations en 2015 à 381 426 déclarations en 2022, soit une progression de 65,5%. De plus, les déclarations de marchandises orientées vers le circuit de contrôle physique (circuit rouge) ont représenté 33,8% en moyenne sur cette période pour un taux de fraude d'environ 3,3% contre 1,7%, 1,0% et 0,4% respectivement pour les circuits jaunes, vert et bleu. En effet, les déclarations de marchandises sont orientées vers quatre (04) circuits de dédouanement en fonction du risque de fraude que présentent ces déclarations. On distingue ainsi le circuit rouge pour les déclarations à risque élevé de fraude et qui nécessitent un contrôle physique du vérificateur en douane, le circuit jaune pour les déclarations à risque modéré nécessitant juste un contrôle documentaire, le circuit vert pour les déclarations sans risque justifiant ainsi une main levée immédiate et enfin le circuit bleu pour les contrôles différés post-dédouanement. Par ailleurs, l'insuffisance des ressources humaines et l'accroissement des déclarations orientées vers les circuits de contrôle, peuvent rallonger les délais de dédouanement des marchandises. La question de l'optimisation du temps nécessaire pour la mainlevée des marchandises trouve ainsi sa légitimité pour rendre compétitives les frontières douanières nationales par la fourniture des biens et services dans les délais et à faible coût.

Cette problématique s'inscrit d'autant plus dans un contexte d'intensification attendue des flux commerciaux avec l'entrée en vigueur de la ZLECAf. À ce titre, Bate et Guedikouma (2023) soutiennent que la mise en œuvre de la ZLECAf entraînera une augmentation significative des échanges entre les pays africains, posant de nouveaux défis aux administrations douanières en matière de contrôle douanier. Dans ce contexte, le recours aux technologies de l'information et de la communication (TIC) et aux techniques plus efficaces d'analyse de méga-données, apparaît comme une voie stratégique pour améliorer la capacité de ciblage des déclarations douanières potentiellement frauduleuses. L'enjeu est de concilier l'objectif de facilitation des échanges avec celui de maximisation des recettes, deux missions cruciales de la douane, sans pour autant opérer des contrôles physiques sur la totalité ou la quasi-totalité des marchandises. Les expériences de certains pays développés (États-Unis (AI for Customs Modernization), Canada (Aurora AI), le Singapour (TradeNet)) et pays en développement comme l'Afrique du Sud, le Maroc, la Tunisie, le Sénégal et la Côte d'Ivoire peuvent être une source d'apprentissage à cet effet pour les autres pays en développement. De plus, avec l'accroissement sans cesse des flux de marchandises au cordon douanier, il est impossible de contrôler toutes les marchandises dans les délais. Ces délais trop longs induisent la perte de compétitivité des ports, entraînant des détournements de trafics vers les frontières où il y'a plus de célérité et engendrant par la même occasion des frais de stockage ou d'entreposage des marchandises aux frontières. Tous

https://doi.org/10.47348/AMTJ/V5/i1a1

ces éléments entrainent une perte considérable des recettes de ports et rappelent donc la nécessité de mettre en place des méthodes modernes d'analyse et de gestion automatisée des risques douaniers.

Ainsi, à travers cette étude, nous voulons vérifier empiriquement que, si des méthodes scientifiques plus élaborées d'analyse de risque étaient utilisées pour cibler automatiquement les déclarations susceptibles de fraude, celles-ci compromettraient-elles les performances de recettes, ou bien au contraire, permettraient-elles de réduire considérablement les contrôles intrusifs au cordon douanier. En d'autres termes, le Machine Learning (ML) et l'analyse miroir sont-ils capables de détecter et de prévoir les infractions douanières de façon efficace au Togo?

En explorant ces questions, l'objectif de cette étude consiste à analyser l'efficacité du ciblage des contrôles douaniers à partir des données issues des sources internes et externes afin d'orienter les contrôles douaniers et réduire considérablement les contrôles intrusifs. Ainsi, cette étude apporte une double contribution à la fois opérationnelle et scientifique. D'un point de vue opérationnel, cette étude, grâce au ciblage plus précis des déclarations potentiellement frauduleuses, permettra de réduire considérablement le nombre de déclarations orientées vers les circuits de contrôle et favoriser un meilleur contrôle de celles-ci par la main d'œuvre disponible. D'un point de vue scientifique, cette étude vient alimenter la littérature sur le débat de l'intérêt que présente les techniques d'apprentissage machine au sein des administrations fiscales et douanières notamment dans les pays en développement.

Le travail est structuré en trois grandes sections. La première section est consacrée à une brève revue de littérature sur l'usage de l'apprentissage machine et l'analyse miroir. La seconde section quant à elle expose la démarche méthodologique, la présentation des données et le protocole de traitement. Enfin l'analyse et l'interprétation des résultats sont présentées dans la troisième section suivie d'une conclusion.

II. Revue de littérature

La littérature en matière de gestion des risques inhérents à la fraude dans les administrations fiscales et douanières est assez très peu fournie en raison de la spécificité du domaine qui tient au secret professionnel (Qinghua et al 2023, Walter D. et al 2020, Tchila, 2020). Toujours est-il qu'en vue de limiter les contrôles intrusifs et faciliter les échanges au cordon douanier, les administrations des douanes ont recours aux techniques de gestion des risques et au contrôle a posteriori fondés sur les données issues de toutes les étapes de la procédure douanière. La gestion des risques permet de prévoir à l'avance les irrégularités pouvant être contenues dans les déclarations en douane et de prendre les mesures appropriées. Elle permet également aux administrations des douanes d'être plus performantes dans l'identification et l'analyse des

https://doi.org/10.47348/AMTJ/V5/i1a1

risques afin d'orienter les contrôles de première ligne et les contrôles après dédouanement. Le processus de gestion des risques comporte l'établissement du contexte de la gestion des risques, l'identification des risques, l'analyse des risques, l'évaluation des risques, le traitement des risques, le suivi et l'évaluation du processus, ainsi que la communication et la consultation (OMD, 2010). Les outils traditionnels de gestion des risques ont été d'abord fondés sur l'appréciation humaine en fonction de l'expérience vécue par les professionnels du métier (Ashtiani M. et al 2021), ensuite le profilage à partir des régressions des séries chronologiques et le scoring effectué avec les modèles de régression linéaire (LOGIT, PROBIT, etc). Cependant ces techniques de gestion des risques fondées sur les méthodes traditionnelles présentent des limites dès lors que les données collectées au cordon douanier sont complexes et volumineuses.

Dans son analyse des systèmes de gestion des risques dans les pays en développement, Laporte (2011) relève le caractère qualitatif du traitement de l'information avec l'analyse de la sélectivité; les critères de sélectivité peu nombreux ainsi que la dualité de l'analyse de ces critères de sélectivité pouvant justifier les taux élevés d'inspections intrusives contre très peu d'infractions relevées. L'auteur dans ces travaux fait recours au modèle de probabilité linéaire, aux modèles LOGIT et PROBIT pour analyser le ciblage des déclarations et parvient à l'efficacité du modèle linéaire par rapport aux modèles non linéaires dans le ciblage des infractions. Il relève toutefois qu'en raison du nombre assez faible des infractions constatées, ces modèles quoique scientifiquement rigoureux ont des difficultés à s'adapter contrairement à l'apprentissage machine. Ainsi, le recours aux techniques mieux élaborées d'apprentissage automatique s'avère nécessaire et adapté en vue d'un meilleur ciblage des déclarations à risque de fraude (Qinghua et al 2023, Mamo, 2013). C'est dans cet ordre d'idée que Mamo (2013) et Bezabeh (2019) relèvent que le processus de gestion des données fondé sur l'apprentissage automatique est régi par les étapes suivantes: l'acquisition des données (sélection du type de données à utiliser), le prétraitement ou l'intégration des données (élimination des valeurs aberrantes, correction des données manquantes), l'exploration des données (choix de la technique d'exploration des données, la classification, le regroupement, la segmentation, etc), la construction et la validation des modèles (rejet des modèles non pertinents et mise en production des modèles efficaces).

Dans le domaine du Machine Learning, les approches d'apprentissage supervisé, non supervisé ainsi que l'apprentissage par renforcement constituent les principales méthodes mobilisées pour l'analyse des données. D'après Abdulalem et al (2022), Sathya et Annama (2013), l'apprentissage supervisé utilise les données étiquetées pour développer des modèles prédictifs alors que l'apprentissage non supervisé est utilisé

https://doi.org/10.47348/AMTJ/V5/i1a1

dans le cadre des données non étiquetées et fait référence à la capacité d'apprentissage et d'organisation des informations sans fournir de signal d'erreur pour évaluer la solution potentielle. Quant à l'apprentissage par renforcement, il permet à la machine de s'adapter en apprenant par essaie et par erreur. Au-delà de l'administration fiscale et douanière et face aux méga-données collectées dans le monde, l'analyse de risques fondée sur l'exploration des données et l'apprentissage automatisé est utilisée dans de nombreux domaines à risques de fraude tels que l'assurance et la finance pour déceler les risques et dans d'autres domaines tels que la télécommunication pour détecter la fraude (Walter et al 2020, Shivakumar & Sanjeev, 2014). En analysant les transactions frauduleuses effectuées par carte de crédit, Alemad (2022) a utilisé les algorithmes KNN (k-nearest neighbors) et SVM (support vecteur machine) de l'apprentissage supervisé et la régression logistique pour parvenir à la conclusion suivant laquelle l'algorithme SVM était le plus adapté dans la détection de la fraude (Abdulalem et al (2022). Par contre en utilisant les techniques d'apprentissage supervisé fondées sur les arbres de décision, les estimateurs bayésiens et l'algorithme KNN, Bezabeh (2019) a trouvé l'algorithme KNN plus performant dans la détection et la prédiction de la fraude douanière. Ashtiani M. et al en 2021 ont passé en revue la littérature existante sur la détection intelligente des fraudes dans les états financiers des entreprises et sont parvenus à la conclusion suivant laquelle les algorithmes supervisés sont plus utilisés que les approches non supervisées comme le clustering en utilisant les données textuelles et audio.

Quant au contrôle a posteriori effectué sur les données après dédouanement, les techniques de l'analyse miroir sont utilisées pour comparer les statistiques douanières collectées au cordon douanier via l'importation à celles déclarées dans la base internationale comme ayant fait l'objet d'exportation. Plusieurs avantages ont été mis à l'actif des statistiques miroirs comme outils permettant d'améliorer l'analyse des risques, l'estimation des pertes potentielles de recettes douanières (Grigoriou et al, 2019). Dans leur analyse de l'importance des données miroirs dans le contrôle du commerce international informel, Carère et Grigoriou (2015) affirment qu'au-delà des raisons structurelles (transbordement, réexportation, etc) ou logistiques, les fausses déclarations délibérées de valeur ou d'espèces dans le but de payer moins de droits de douane peuvent expliquer les écarts miroirs. L'analyse miroir vient compléter les outils d'analyse de risque de première ligne lorsque ceux-ci sont obsolètes surtout dans un contexte de facilitation des échanges et d'aléa moral. Elle offre alors la possibilité de détecter la fraude douanière qui a échappé au contrôle de première ligne (Cariolle et al, 2017).

https://doi.org/10.47348/AMTJ/V5/i1a1

Bien que le Machine Learning et l'analyse miroir reposent sur des sources de données et des techniques d'analyse distinctes, ces deux approches s'avèrent complémentaires dans la lutte contre la fraude douanière. Le Machine Learning, en exploitant les données disponibles au cordon douanier, permet de cibler en temps réel les déclarations à risque et d'orienter les contrôles de première ligne. En revanche, l'analyse miroir, en comparant les données commerciales bilatérales, constitue un outil efficace d'investigation a posteriori, en appui aux équipes en charge des contrôles différés.

III. Démarche méthodologique

III.1 Cadre théorique

D'un point de vue théorique, la lutte contre la fraude douanière repose sur le modèle d'Allingham et Sandmo (1972), selon lequel les agents économiques font un arbitrage entre le gain attendu d'une fraude et le coût espéré en cas de détection. Dans ce contexte, améliorer la capacité de détection constitue un levier essentiel pour dissuader les comportements frauduleux. L'approche d'analyse des risques, adoptée par les administrations fiscales et douanières, vise à identifier les opérations les plus suspectes afin d'optimiser les ressources de contrôle et de concilier l'efficacité administrative et la fluidité des échanges.

Les avancées en science des données permettent désormais d'outiller cette démarche grâce aux algorithmes de Machine Learning supervisé, capables d'apprendre à partir d'exemples historiques pour prédire la probabilité d'une fraude. Ces techniques offrent un avantage comparatif en matière de détection des non-conformités complexes et non linéaires, souvent invisibles via les méthodes classiques. L'ensemble de ces outils s'inscrit dans une logique de pilotage fondé sur les données (data-driven decision-making), en phase avec les exigences contemporaines des administrations douanières confrontées à une croissance exponentielle des flux commerciaux.

III.2 Méthode d'analyse

Dans le cadre de cette étude, deux grandes approches méthodologiques sont mobilisées pour détecter les cas potentiels de fraude douanière. Il s'agit des algorithmes de Machine Learning supervisé et l'analyse miroir utilisé comme méthode comparative d'analyse des écarts commerciaux.

https://doi.org/10.47348/AMTJ/V5/i1a1

III.2.1 Spécification du modèle

Rappelons que l'apprentissage machine est une démarche qui consiste au-delà de la collecte et la gestion des données à l'analyse, la prédiction et l'utilisation des algorithmes qui améliorent automatiquement leurs performances grâce à l'apprentissage (Mamo, 2013). Ces algorithmes optimisent les opérations statistiques en analysant les données d'entrée pour réaliser des prédictions dans une plage acceptable. Selon Mamo (2013), les méthodes de Machine Learning peuvent être divisées en trois grandes catégories à savoir: les méthodes d'apprentissage supervisé, les méthodes d'apprentissage non supervisé et les méthodes d'apprentissage semi-supervisé.

Les méthodes d'apprentissage supervisé sont une classe d'algorithme d'apprentissage automatique qui utilisent des données étiquetées pour entrainer un modèle à prédire des résultats futurs. Les méthodes d'apprentissage non supervisé sont quant à elles une classe d'algorithmes d'apprentissage automatique qui sont utilisées pour découvrir des motifs ou des structures intrinsèques dans les données et cela sans que les données ne soient préalablement étiquetées. Contrairement aux méthodes d'apprentissage supervisé qui nécessitent l'étiquetage des données pour l'entrainement, les méthodes d'apprentissage non supervisé parcourent les données sans aucune indication sur les résultats attendus. Les méthodes d'apprentissage semi-supervisé combinent quant à elles les méthodes d'apprentissage supervisé avec les méthodes d'apprentissage non supervisé. Ces méthodes sont utilisées lorsque les données étiquetées sont difficiles à obtenir du fait de leur coût et qu'au contraire les données non étiquetées sont disponibles en abondance.

Dans le cadre de cette étude, l'approche adoptée est celle des méthodes d'apprentissage supervisé. Le choix de cette technique se justifie par le fait que dans la base de données SYDONIA, l'information sur le statut de déclaration (suivant qu'elle soit réajustée ou non ou suivant qu'elle soit classée dans un circuit de contrôle ou non) est disponible. Ces méthodes sont très pratiques et permettent une meilleure compréhension des caractéristiques qui influent sur la prédiction.

Ces modèles ont été entrainés sur un échantillon labellisé à partir d'exemples historiques puis testés sur des observations indépendantes. Le principe repose sur l'entraînement d'un algorithme à partir d'un ensemble de données étiquetées, où chaque observation est caractérisée par un ensemble de variables explicatives et une variable cible binaire indiquant la présence ou l'absence de fraude.

Soit: $Y_i \in \{0,1\}$ la variable cible $Y_i = 1$ indique une opération frauduleuse, et $Y_i = 0$ une opération régulière.

$X_i = \left(x_{i1}, x_{i2}, \ldots, x_{ik}\right)$ le vecteur des caractéristiques observées pour l'observation *i*, telles que l'origine de la marchandise, la masse nette, la valeur, le bureau de dédouanement, l'espèce tarifaire, etc.

https://doi.org/10.47348/AMTJ/V5/i1a1

Le modèle supervisé prend une fonction f:X → Y, telle que: $\widehat{Y_i} = f(X_i)$ où $\widehat{Y_i}$ est la prédiction de la probabilité de fraude associée à l'observation *i*. Plusieurs algorithmes de classification peuvent être utilisés pour estimer cette fonction *f* notamment la régression logistique, les arbres de décision, le random Forest, le XGBoost et le réseau de neurones.

III.2.2 Technique d'apprentissage

Dans les algorithmes d'apprentissage automatique supervisé, un ensemble de données d'entraînement étiqueté est d'abord utilisé pour entraîner l'algorithme sous-jacent en attribuant un score de risque (Sathya et Annama, 2013). Cet algorithme entraîné est ensuite alimenté par l'ensemble de données de test non étiquetées pour les classer en groupes similaires. Les algorithmes d'apprentissage supervisé conviennent bien à deux types de problèmes: les problèmes de classification et des problèmes de régression.

Afin de construire le modèle de prédiction de la fraude douanière, l'intérêt est porté sur la méthode de classification dans laquelle la variable de sortie sous-jacente est discrète. Cette variable sera classée en deux (02) groupes ou catégories. Le premier groupe sera le groupe des fraudeurs et le second groupe sera le groupe des non-fraudeurs. Plusieurs algorithmes d'apprentissage supervisé peuvent être utilisés pour prédire l'état d'une déclaration au cordon douanier, cependant afin de faciliter l'interprétation des résultats, la robustesse de chaque algorithme, et la complémentarité des différents algorithmes, nous faisons un usage simultané des modèles suivants: le modèle de régression logistique (LR), le modèle d'arbre de décision (DT ou decision tree), le modèle XGBoost, l'algorithme des réseaux de neurones et la méthode du Random Forest (Forêt Aléatoire).

Le modèle d'arbre de décision

Un arbre de décision modélise la logique de décision, en testant et en classant les éléments de données dans une structure arborescente (Mamo, 2013). Chaque nœud d'un arbre de décision représente une fonctionnalité dans une instance à classer, et chaque branche représente une valeur que le nœud peut prendre. Les instances sont classées en commençant par le nœud racine et triées en fonction de leurs valeurs de fonctionnalités.

Les instances sont classées à partir du nœud racine et suivent les branches en fonction des valeurs des caractéristiques jusqu'aux nœuds feuilles, ce qui correspondent aux résultats de la décision (Shahadat et al 2019). Par ailleurs, les arbres de décision sont considérés comme résistants au bruit car leurs stratégies d'élagage évitent le sur-ajustement des données en général et des données bruyantes en particulier. Ils sont

https://doi.org/10.47348/AMTJ/V5/i1a1

faciles à interpréter et à visualiser même s'ils sont sensibles à de petites variations dans les données qui entrainent très souvent des arbres très différents.

L'algorithme de réseau de neurones

C'est un modèle mathématique et informatique qui est basé sur les réseaux neuronaux biologiques, c'est-à-dire qu'il imite la façon dont le cerveau humain apprend. Dans ce modèle, les connaissances sont acquises par le réseau grâce à un processus d'apprentissage et les nœuds des neurones sont utilisés pour stocker des informations. Le réseau neuronal peut être représenté comme un graphe dans lequel chaque nœud exécute une fonction de transfert de la forme

$$Y_i = \left(f_i \sum_{j=1}^{n} W_{ij} X_j - Q_i \right) \quad (1)$$

Y_i est la sortie du nœud
X_i est la j ième entrée du nœud
W_{ij} est le poids de connexion entre les noeuds i et j

L'algorithme des réseaux de neurones est capable de modéliser les relations complexes et non linéaires dans les données. Il est très flexible contrairement aux modèles d'arbre de décision et est capable d'apprendre de grandes quantités de données. Cependant, ce type d'algorithme nécessite un long temps d'entraînement et exige une grande quantité de données et une forte puissance de calcul.

Il faut noter qu'il y'a deux types de réseaux neuronaux notamment le réseau à avance direct pour lequel l'information se déplace dans une seule direction vers l'avant entre les nœuds d'entrée et de sortie en passant par les nœuds cachés s'il y a lieu et les réseaux récurrents qui sont des modèles à flux de données bidirectionnels.

La méthode du classificateur XGBoost

Le classificateur XGBoost ou Extreme Gradient Boosting est une implémentation efficace de l'algorithme d'amplification de gradient consistant à bâtir une prédiction fiable en agrégeant les réponses d'apprenants de base. Le principe de ce modèle consiste à combiner les résultats d'un ensemble de modèles plus simple et plus faibles afin de fournir une meilleure prédiction. Il s'agit d'une méthode d'agrégation de modèles: au lieu d'utiliser un seul modèle, l'algorithme va en utiliser plusieurs qui serons ensuite combinés pour obtenir un seul résultat.

C'est un modèle très performant et qui est souvent considéré comme l'un des meilleurs modèles dans les compétitions de Machine Learning. Il est capable de traiter les valeurs manquantes et permet d'éviter le sur-

https://doi.org/10.47348/AMTJ/V5/i1a1

ajustement. Cependant, il est très difficile à interpréter en raison de la nature du modèle d'ensemble.

La régression logistique

La régression logistique, plus précisément le modèle Logit est une technique permettant d'ajuster une surface de régression à des données lorsque la variable dépendante est dichotomique. Elle est utilisée pour des études ayant pour but de vérifier si des variables indépendantes peuvent prédire une variable dépendante dichotomique. Il s'agit en fait de connaître les facteurs associés à un phénomène en élaborant un modèle de prédiction.

Le modèle Logit repose sur la fonction de lien logistique (ou sigmoïde), qui transforme la variable de réponse en probabilités comprises entre 0 et 1. Contrairement aux modèles linéaires classiques, la régression logistique n'exige pas que les prédicteurs soient distribués normalement, ni qu'ils présentent une variance égale entre les groupes. Son efficacité tient de sa capacité à isoler les données appartenant à différentes classes binaires (Hala et Nojood, 2020).

Tout comme le modèle d'arbre de décision, le modèle de régression logistique est très simple à interpréter et à implémenter en plus d'être robuste même avec un petit de jeu de données. Cependant, le modèle de régression logistique ne capture pas les relations linéaires entre la variable cible et les variables dépendantes.

Le Random Forest

Le Random Forest est composé de plusieurs arbres de décision, entrainés de manière indépendante sur des sous-ensembles du jeu de données d'apprentissage (méthode de bagging). Chacun produit une estimation, et c'est la combinaison des résultats qui va donner la prédiction finale qui se traduit par une variance réduite. Chaque modèle est distribué de façon aléatoire en sous-ensembles d›arbres décisionnels. Cet algorithme présente l›avantage d›ajouter un caractère aléatoire au modèle en trouvant la meilleure caractéristique parmi un sous-ensemble aléatoire de caractéristiques et réduit la variance des prévisions d'un seul arbre de décision, améliorant ainsi les performances.

Le Random Forest, tout comme le modèle de réseaux de neurones est capable de gérer les jeux de données de grandes tailles avec de nombreuses variables. Il est plus robuste aux bruits que les modèles d'arbre de décision simples. Ce type de modèle tout comme le modèle de réseaux de neurones, est lent à entraîner lorsque le nombre d'arbre est important et nécessite plus de ressources en termes de mémoire et de calcul.

https://doi.org/10.47348/AMTJ/V5/i1a1

L'analyse miroir

La démarche méthodologique de l'analyse miroir consiste pour un produit donné, à comparer les exportations déclarées par le pays exportateur et les importations reportées par le pays importateur, destinataire du produit, afin de détecter notamment les écarts de quantité, de poids ou de valeur. Afin de calculer ces écarts miroir, nous nous basons sur le modèle de Ndikumana et Boyce (2021) qui dégage les écarts commerciaux à partir de l'équation suivante:

$$D X_k^t = \sum_{j=1,i=1}^{JI} (M_{ji,t}^k - \beta X_{ji,t}^k) \quad (2)$$

DX représente l'écart commercial entre les importations (M) et les exportations (X), en poids ou en valeur des marchandises, objet du commerce international. M représente les importations déclarées au cordon douanier par le pays importateur dans le système de dédouanement national, X représente les exportations déclarées par les partenaires commerciaux dans la base internationale COMTRADE. Enfin, β représente le coefficient de corrélation lié à la variable exportation.

Cette équation permet de dégager les écarts commerciaux entre les importations effectivement reçues au Togo et les exportations déclarées par les partenaires commerciaux et permet explicitement de comparer les données des mêmes flux de marchandises issues de la base de données internationale (COMTRADE) extraite de WITS, un site de la banque mondiale et de la base nationale (SYDONIA WORLD) qui enregistre les informations sur les échanges commerciaux au cordon douanier, notamment les importations effectuées par le Togo.

III.3 Sources de données et mesure

Les données utilisées pour l'apprentissage machine portent sur les déclarations en douane du Togo sur les années 2017 à 2022 et sont issues du système douanier automatisé (SYDONIA). Le jeu de données ainsi constitué comporte deux millions vingt-quatre mille deux cents vingt-huit (2 048 228) observations. Chaque observation correspond à une déclaration de douane enregistré par un opérateur par l'intermédiaire de son commissionnaire en douane. Un protocole de traitement a été appliqué aussi bien à la variable cible dépendante qu'aux variables prédictives indépendantes. Au total, dix (10) variables ont été retenues, dont la variable cible de prédiction.

La variable cible est le statut de la déclaration par rapport à la fraude. Elle indique si une déclaration a fait l'objet d'un réajustement du montant des droits et taxes à payer, à la suite de la constatation d'une fraude avérée par le vérificateur en douane. C'est une variable catégorielle qui ne prend que deux modalités (fraude ou non fraude).

https://doi.org/10.47348/AMTJ/V5/i1a1

Les autres variables explicatives sont les suivantes: le statut de renseignement du certificat de visite, indiquant si celui-ci a été renseigné ou non; la conformité des documents, qui renseigne sur l'exactitude des éléments figurant dans la déclaration et constitue un indicateur de risque; le circuit de dédouanement, en lien avec le niveau de risque de fraude présenté par la marchandise importée, reparti en quatre types (rouge, vert, bleu et jaune). Les circuits rouge et jaune sont associés au contrôle physique et documentaire tandis que les circuits bleu et vert correspondent à des circuits hors contrôle. Sont également pris en compte: la localisation du bureau de dédouanement choisi par l'opérateur, l'espèce tarifaire correspondant à la dénomination de la marchandise dans le tarif extérieur commun (TEC), et le régime douanier définissant le statut de la marchandise (importation, exportation, transit, admission temporaire, etc). Par ailleurs, l'origine de la marchandise c'est-à-dire le pays d'origine est intégré, car elle influence la valeur de la marchandise et le régime de taxation applicable. Enfin deux variables quantitatives complètent les données: la valeur déclarée de la marchandise (valeur caf de la déclaration) et la masse nette de l'article contenu dans la déclaration, cette dernière permettant d'évaluer l'homogénéité entre les articles et de détecter des anomalies éventuelles.

Le choix de ses variables s'est opéré à la suite des techniques statistiques telles que les tests d'indépendances statistiques (test de khi-deux) et les tests de colinéarité. Le traitement de la variable cible (fraude) a exigé que lorsque les droits et taxes compromis d'une déclaration sont supérieurs à zéro, celle-ci est considérée comme frauduleuse, dans le cas contraire, on considère que la déclaration n'est pas frauduleuse. Par ailleurs le traitement des autres variables a consisté à la correction des données manquantes, des données aberrantes ainsi que la correction des modalités. Les variables catégorielles ont fait l'objet de l'encodage suivant la méthode one-hot tandis que les variables numériques ont subi une normalisation des grandeurs. Après ces étapes, le jeu de données est désormais constitué d'un million cinq cent quatre-vingt-dix-sept mille deux cent quatre-vingt-huit (1 597 288) déclarations et de soixante-trois (63) variables. Pour rappel, dans le cadre de ce travail, la fraude a lieu lorsqu'il y'a une différence entre les droits et taxes liquidés par l'opérateur économique et ceux liquidés par les agents aux cordons douaniers. Cette définition de la fraude s'est matérialisée par un déséquilibre des classes puisque seuls 3,3% des déclarations étaient étiquetés comme frauduleux contre 96,7% des déclarations qui ne l'étaient pas. Pour assurer l'apprentissage, l'échantillon d'apprentissage des données a été rééquilibré par la méthode de sur-échantillonnage appelée Technique de suréchantillonnage synthétique de la classe minoritaire (SMOTE). C'est une technique qui consiste en effet à générer de nouvelles déclarations de la classe minoritaire qui ressemblent aux autres déclarations de la même

https://doi.org/10.47348/AMTJ/V5/i1a1

classe sans être identiques. Avant d'appliquer les modèles d'apprentissage supervisés décrits plus haut, le jeu de données a été subdivisé en base d'entrainement (80%) en base test (10%) et en base de validation (10%). Cette subdivision a été faite en se basant sur les travaux déjà réalisés dans le cadre de la détection de la fraude douanière.

En ce qui concerne l'analyse miroir, en complément des données nationales issues de SYDOONIA pour l'année 2020, les données du commerce extérieur provenant des pays partenaires du Togo sont également utilisées. Ces données extraites de la base COMTRADE, portent principalement sur l'année 2020. Plusieurs opérations de traitement ont été nécessaires afin de garantir la comparabilité et la cohérence des données issues des deux sources. Premièrement, les données issues de COMTRADE ont été converties pour adopter le format utilisé dans SYDONIA. Ensuite, une sélection rigoureuse des régimes douaniers pertinents a été effectuée, en tenant compte uniquement des régimes appliqués aux marchandises à destination du territoire togolais. Cette étape est cruciale pour éviter la double comptabilisation de certaines informations.

Par ailleurs, les données relatives au pays qui ne figurent pas dans COMTRADE 2020 ont été écartées, tout comme celles concernant les marchandises issues de la Zone Franche du Togo. Les données SYDONIA ont été agrégées au niveau du code SH6 afin d'assurer une granularité cohérente avec celle des données COMTRADE, Un rapprochement des deux bases a ensuite été réalisé, suivi du calcul des écarts observés entre les données déclarées par le Togo et celles rapportées par ses partenaires commerciaux. Enfin, il convient de noter que seuls les régimes directs ont été pris en compte dans cette analyse, y compris les régimes suspensifs. Les régimes d'exportation quant à eux, ont été systématiquement exclus de l'étude.

IV. Résultats et discussions

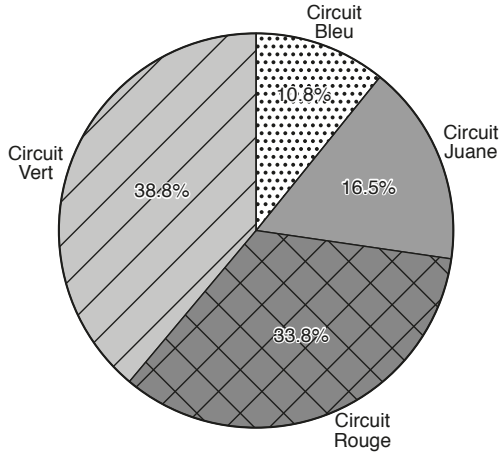
Cette section centrée sur l'analyse des résultats, expose dans un premier temps, les résultats du ML et dans un second temps, les résultats de l'analyse miroir. Mais bien avant, il convient de faire une présentation de quelques résultats de l'analyse exploratoire des données.

IV.1 Analyse exploratoire des données

Comme souligné plus haut, sur la période 2017-2022, plus de deux millions de déclarations ont été enregistrées. Parmi celles-ci et conformément au graphique 1, les déclarations orientées au circuit de contrôle physique (circuit rouge) occupent en moyenne une part importante avec une proportion de 33,8% des déclarations. La part des déclarations au circuit de contrôle documentaire (circuit jaune) s'élève à 16,5% tandis que la proportion des déclarations hors circuits de

https://doi.org/10.47348/AMTJ/V5/i1a1

contrôle est respectivement de 38,8% pour le circuit vert et de 10,8% pour le circuit bleu. Malgré la part élevée des déclarations au circuit vert, il convient de noter que les déclarations au circuit rouge ont été plus importantes en 2017, 2018 et 2019. Cette forme de transition qui s'est amorcée entre ces deux circuits de dédouanement s'explique par la volonté de la douane togolaise de répondre aux exigences sur la facilitation des échanges. Toutefois, malgré cette transition, le nombre de déclarations orientées vers le circuit de contrôle physique demeure encore important.

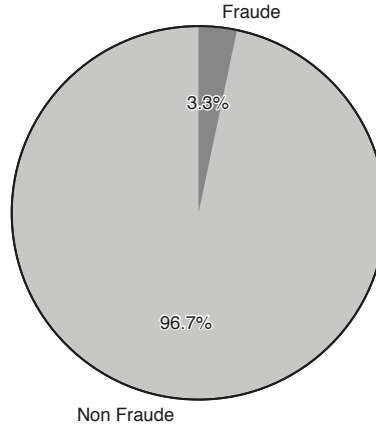


Graphique 1: *Répartition des déclarations selon le circuit de dédouanement*

Source: *Auteurs à partir des données de SYDONIA WORLD de 2017 à 2022*

Il faut noter que la fraude en matière douanière n'est enregistrée que lorsqu'elle est constatée par une minoration des droits et taxes de port. Suivant cette constatation, les cas de fraude constatés dans le jeu de données sont très minoritaires avec seulement 3,3% des déclrations frauduleuses contre 96,7% des déclarations potentiellement non frauduleuses (voir graphique 2).

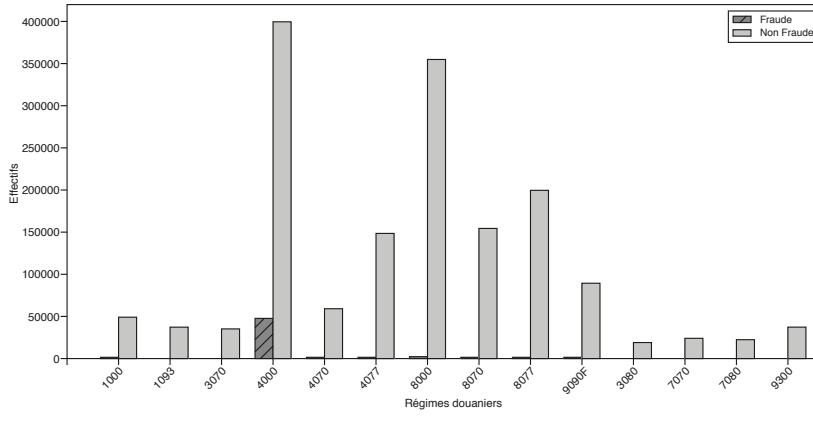
https://doi.org/10.47348/AMTJ/V5/i1a1



Graphique 2: *Statut des déclarations en douane*

Source: *Auteurs à partir des données de SYDONIA WORLD de 2017 à 2022*

L'analyse de la fraude suivant le régime douanier, montre que la fraude est beaucoup plus présente pour les régimes de mise à la consommation directe (2,7%) ou de mise à la consommation après transit (0,2%). Des taux élevés de fraude sont également présents dans les déclarations faisant objet de transit.



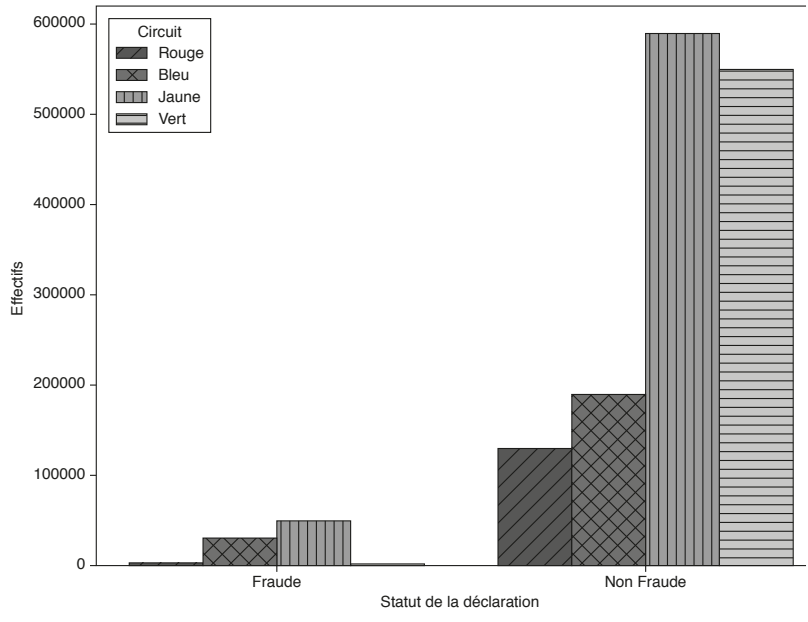
Graphique 3: *Statut des déclarations par régime*

Source: *Auteurs à partir des données de SYDONIA WORLD de 2017 à 2022*

Outre les analyses portant sur les régimes, on constate à travers l'analyse du graphique 4 que les déclarations orientées en ciruit rouge sont assez representées dans le groupe des déclarations frauduleuses. Il faut quand même souligner que le graphique 4 ci-dessous indique également une grande proportion de déclarations orientées en circuit rouge qui ne

https://doi.org/10.47348/AMTJ/V5/i1a1

sont pas frauduleuses. Ceci indique que le système d'orientation des marchandises en circuit pourrait être défaillant et doit donc faire objet de révision.



Graphique 4: *Statut des déclarations par type de circuit*

Source: *Auteurs à partir des données de SYDONIA WORLD de 2017 à 2022*

IV.2 Résultats de l'apprentissage machine

Les résultats de l'apprentissage machine sont analysés à travers plusieurs métriques d'évaluation qui traduisent la performance des modèles suivant des critères bien définis. Dans le cadre de ce travail, quatre métriques d'évaluation ont été retenues à savoir le score de précision, le score F1, le score de rappel et l'AUC. Ces métriques sont calculées sur la base de la matrice de confusion.

IV.2.1 La matrice de confusion

La matrice de confusion est une matrice qui présente de façon résumée les résultats de prédiction. Elle compare en effet, les données réelles (celles de la base d'origine) de la variable cible avec celles produites par le modèle. Cette matrice est utilisée pour calculer les métriques telles que la précision, le score F1 et le score de rappel (Murorunkwere et al, 2023). Ainsi, on pose:

https://doi.org/10.47348/AMTJ/V5/i1a1

TP, les vrais positifs: ils indiquent le cas où les prédictions et les valeurs réelles sont effectivement identiques (positives).
TN, les vrais négatifs: ils indiquent le cas où les prédictions et les valeurs réelles sont négatives.
FP, les faux positifs: ils indiquent une prédiction positive par rapport aux valeurs réelles qui sont négatives
FN, les faux négatifs: ils indiquent une prédiction négative alors que les valeurs réelles sont positives. Les métriques sont alors définies suivant les formules suivantes:

$$Précision = \frac{TP}{TP + FP} \quad (3)$$

La précision mesure en effet la proportion des prédictions positives qui sont effectivement correctes. Dans le cas de la prédiction de la fraude douanière, une précision élevée signifie que lorsque le modèle prédit une fraude, elle est souvent correcte. Cette metrique est très importante car elle identifier une entreprise comme frauduleuse alors qu'elle ne l'est pas peut engendrer des conséquences importantes pour l'administration.

$$Rappel = \frac{TP}{TP + FN} \quad (4)$$

Le Rappel mesure la capacité du modèle à détecter les fraudes réelles, ce qui est très important si l'objectif est de ne pas manquer les fraudes potentielles.

$$Score\ F1 = \frac{Rappel * Précsion * 2}{Rappel + Précision} \quad (5)$$

Le Score F1 correspond en effet à la moyenne harmonique entre la précision et le recall. Cette une métrique qui prend en compte les deux autres métriques afin de donner une vue d'ensemble de l'efficacité du modèle. Dans le cadre de la prédiction de fraude au cordon douanier, un score élevé montre que le modèle parvient à maintenir un bon équilibre entre la détection des cas de fraudes réelles et la réduction des fausses alertes.

Enfin, la métrique AUC-ROC est utilisée pour comparer différents modèles de classification. Elle permet très souvent de déterminer le modèle le plus robuste parmis différents modèles d'apprentissage. Une valeur d'AUC élevé (proche de 1) signifie que le modèle permet de différencier correctement les différentes classes et donc par conséquent signifie que le modèle est puissant (Bowers et Zhou, 2019).

IV.2.2 Analyse des résultats

Après avoir divisé notre ensemble de données en base d'entrainement (80%), en base test (10%) et en base de validation (10%), et après avoir resolu le problème de déséquilibre des classes par la méthode de sur-

https://doi.org/10.47348/AMTJ/V5/i1a1

échantillonnage (SMOTE), nous avons obtenu les résultats tels que présentés dans le tableau ci-dessous.

Tableau 1: *Comparaison des métriques d'évaluation pour différents modèles*

Modèles	F1-score	Recall	Précision
Regression logistique	0,78	0,76	0,80
Arbres de décision	0,72	0,72	0,71
XgBoost	0,79	0,75	0,83
Forêt aléatoire	0,77	0,74	0,81
Réseaux de neurones	0,79	0,76	0,82

Source: *Auteurs à partir des données de SYDONIA WORLD*

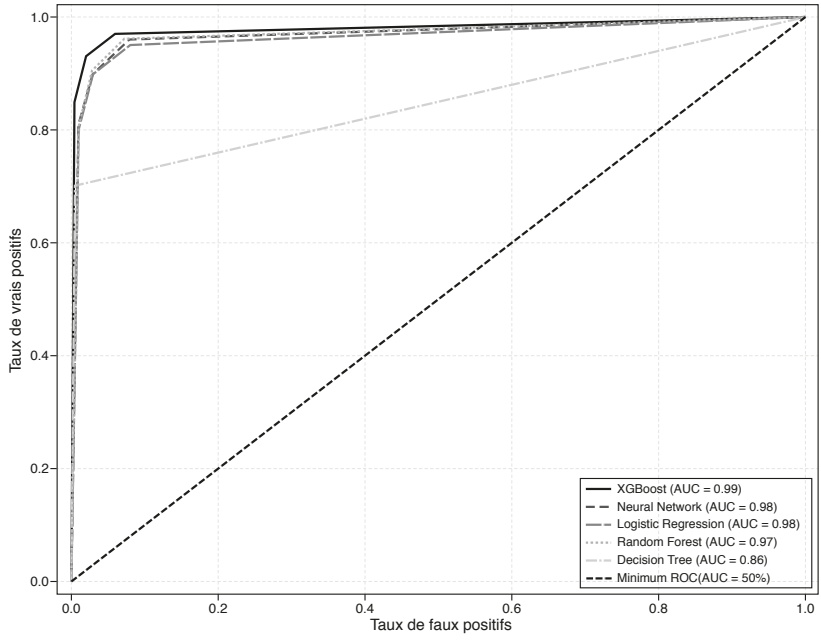
A l'analyse du tableau ci-dessus, on constate que pour ce qui est du f1-score, le XGBoost et le modèle de réseaux de neurones ont les scores les plus élévés (0,79). Cela suggère donc que ces modèles ont une bonne balance entre précision et rappel. Pour rappel, le f1-score combine la précision et le rappel en un seul nombre minimisant par la même occasion à la fois ou les faux positifs et les faux négatifs.

Le recall mesure la proportion de vrai positifs (c'est à dire les cas de fraude) parmi tous les exemples réellement positif. Le modèle de regression logistique et le modèle de réseaux de neurones ont le rappel le plus élévé (0,76) ce qui signifie qu'ils sont meilleurs pour détecter la majorité des déclarations positives (frauduleuses).

La précision mesure la proportion de prédictions positives correctes parmi toutes les prédictions positives. En d'autres termes la précision mesure la capacité du modèle à ne pas classer à tort les déclarations négatives comme positives. Dans le cas de notre étude, les modèles XGBoost a la précision la plus élévée, ce qui indique qu'il est meilleur pour éviter les faux positifs.

Le graphiue 5 ci-dessous compare l'AUC-ROC des modèles étudiés et montre le modèle le plus robuste parmi les classificateurs utilisés dans cette étude de recherche. Selon les valeurs de l'AUC-ROC présentés sur le graphique 6 ci-dessous, tous les modèles semblent performants mais le modèle XGBoost (0,99) se démarque comme le meilleur en terme de capacité à détecter les déclarations frauduleuses, suivi de près par le modèle de réseaux de neurones et de regression logistique (0,98).

https://doi.org/10.47348/AMTJ/V5/i1a1



Graphique 5: *Comparaison des modèles à partir des courbes AUC-ROC*

Source: *Auteurs à partir des données de SYDONIA WORLD*

IV.3 Résultats de l'analyse miroir

Comme souligné dans la méthodologie, l'analyse miroir qui survient a posteriori vient compléter les résultats de l'apprentissage machine en première ligne. Elle permet ainsi de détecter les nouveaux schémas de fraudes et les cas de fraudes qui n'ont pu être détectés en première ligne. Celles-ci permettront d'alimenter la base de données futur de l'apprentissage machine pour améliorer progressivement le ciblage des infractions en première ligne. Les résultats de l'analyse miroir ont permis de dégager trois situations possibles pour de l'écart commercial DX dans l'équation (2) suivant les flux de marchandises:

- Les flux avec correspondance :il s'agit des exportations des pays partenaires qui sont effectivement reportées dans Sydonia pour les mêmes espèces déclarées à l'importation au Togo, Les poids et valeur sont relativement similaires;
- Les importations orphelines: elles concernent les flux d'importations déclarés dans SYDONIA WORLD mais non renseigné dans WITS;
- Les exportations perdues: c'est le cas des exportations enregistrées par les pays partenaires dans WITS sans qu'il existe de déclarations d'importations enregistrées dans SYDONIA WORLD.

https://doi.org/10.47348/AMTJ/V5/i1a1

Le but de l'analyse consiste à dégager les marchandises prioritaires objets de fraudes potentielles, ce qui permettra d'identifier des entreprises dans lesquelles un contrôle a posteriori pourrait être effectué et permettrait de procéder à des recouvrements conséquents des droits éludés.

Tableau 2: *Marchandises prioritaires objets de fraude potentielle*

SH6	Libellé chapitre	(M-X) MASSE (En millions)	(M/X) MASSE	(M-X) VAL (En millions)	(M/X) VAL	DROITS COMPROMIS (En millions)	Type de fraude
520852	Tissus de coton, contenant moins de 85% en poids de coton, d'un poids n'excédant pas 200 g/m2, imprimés, obtenus par procédé d'impression à base de cire (WAX)	-51,897	1,666	-143,563	2,446	-2952,517	Glissement tarifaire
851714	Autres téléphones pour réseaux cellulaires ou autres réseaux sans fils	2,639	0,465	-157,412	4,738	-1340,06	Sous-évaluation
630900	Articles de friperie	-41,064	2,096	-17,834	2,394	-279,761	Glissement tarifaire
540754	Autres tissus, contenant moins de 85% en poids de filaments de polyester textures imprimés	-16,753	5,919	-30,575	10,37	-226,255	Glissement tarifaire
100630	Riz en paille (riz paddy) semi blanchi ou blanchi en emballage de plus de 5kg ou en vrac	-351,582	0,152	-61,099	9,464	-215,797	Glissement tarifaire
.....	.....	.....		.....		.....	.....

Source: *Auteurs à partir des données de SYDONIA WORLD et COMTRADE*

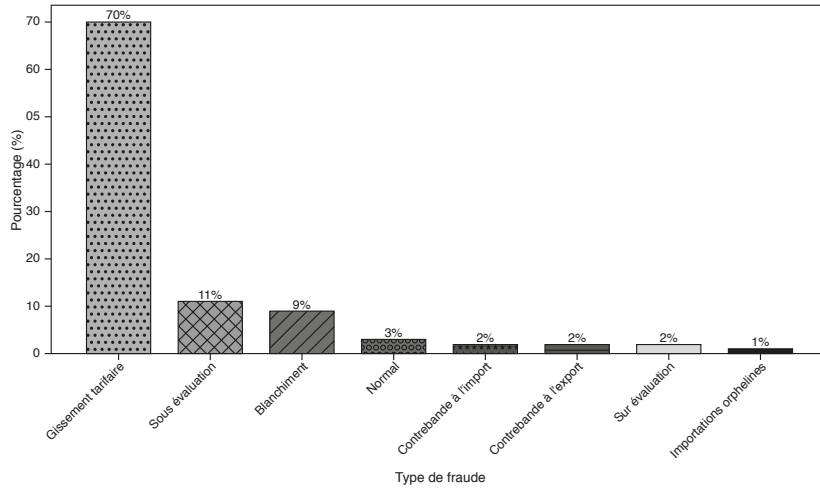
Les marchandises susceptibles de fraude, pour lesquelles l'administration pourrait dégager des droits éludés élevés dans les contrôles a posteriori sont: les tissus coton des chapitres 52 et 54, les téléphones portables du chapitre 85, les friperies du chapitre 63 et le riz du chapitre 10.

https://doi.org/10.47348/AMTJ/V5/i1a1

Ces espèces ont été retenues du fait de leurs écarts significatifs et des possibilités de glissements tarifaires, de la minoration de poids et/ou de valeur dont elles peuvent faire l'objet.

Ces positions représentent 20% des marchandises pouvant permettre à la douane de recouvrer 80% des droits et taxes compromis (méthode revisitée). Ceci permet au contrôle a posteriori de cibler les entreprises à contrôler, augmentant ainsi le rendement et l'efficacité du contrôle.

Le graphique ci-dessous présente la présomption de fraude dans les positions prioritaires.



Graphique 6: *Présomption de fraude dans les positions prioritaires*

Source: *Auteurs à partir des données de SYDONIA WORLD et COMTRADE*

Ce graphique montre que le glissement tarifaire a été la fraude la plus courante (70%) et a représenté 1 752 déclarations en douane uniques dans la procédure de dédouanent sur les présomptions de fraudes potentielles. Il est suivi de la sous-évaluation qui est portée à 11%. La forte proportion du glissement tarifaire se justifie par le fait que pour des quotités élevées (DD=20% ou 35%), les opérateurs économiques sont incités à procéder à des fausses déclarations d'espèces en déclarant de faibles quotités (DD=0% 5% ou 10%) dans le but de payer moins de droits et taxes de douane.

Les statistiques miroir sont plus utiles lorsqu'elles sont complétées par des enquêtes et d'autres méthodes d'investigation. Il est important de recouper et d'enquêter de manière approfondie avec d'autres sources d'information pour confirmer la fraude.

https://doi.org/10.47348/AMTJ/V5/i1a1

V. Conclusion

Dans le cadre de ce travail, l'objectif a été de vérifier empiriquement à partir des données issues du système informatique douanier au Togo, l'efficacité des méthodes sophistiquées d'analyse de risque afin d'orienter les contrôles de première et de deuxième ligne. Pour ce faire, nous avons fait usage d'analyse de l'apprentissage machine et des données miroir. Les résultats des expérimentations montrent que certains algorithmes en particulier XGBoost, les réseaux de neurones et, dans une moindre mesure, la régression logistique, présentent des performances élevées, avec des scores F1 supérieurs à 0,78 et des courbes AUC-ROC proches de 1, indiquant une forte capacité de détection des fraudes tout en limitant les erreurs de classement. Cette robustesse des performances confirme que ces approches peuvent efficacement améliorer le ciblage des contrôles douaniers, tant en première qu'en seconde ligne, tout en respectant l'équilibre nécessaire entre l'efficacité des contrôles et la fluidité du commerce.

Au regard des résultats obtenus, les recommandations suivantes peuvent être formulées à l'endroit de l'administration douanière togolaise:

- Mettre en place une base de données et un système d'information complet sur les différents types de fraudes. Le présent travail a permis de constater une grande quantité de données aberrantes et manquantes. La mise en place d'un système d'information complet et dynamique permettra d'améliorer efficacement le ciblage des infractions douanières;
- Mettre en place un modèle d'apprentissage machine pour améliorer l'efficacité des contrôles et la facilitation des échanges. En effet, conformément aux métriques d'évaluation, le protocole d'apprentissage machine mis en place pour cette étude, permet de cibler efficacement la fraude comparer à la situation actuelle. On note par exemple qu'environ 99% des déclarations qui ont été prédites comme frauduleuses le sont effectivement;
- Orienter les déclarations au circuit de contrôle pour lequel elles doivent effectivement être orientées et exercer à cet effet le contrôle correspondant au dit circuit. Cela évite d'opérer des contrôles intrusifs sur les déclarations orientées vers des circuits hors contrôles;
- Redynamiser l'analyse des données miroir. Actualiser régulièrement les critères de risque en fonction des cas de fraude détectés.

La prise en compte de ces suggestions permettra sans doute d'améliorer la prédiction et la détection de la fraude. Toutefois, il convient préalablement de former un personnel dédié et stable sur les méthodes

https://doi.org/10.47348/AMTJ/V5/i1a1

modernes d'analyse de donnée pour une utilisation efficace de l'apprentissage machine et l'analyse miroir.

Références

Allingham, M., & Sandmo, A. (1972). 'Income tax evasion: a theoretical analysis'. *Journal of Public Economics*, vol.1, pp.323–338.

Abdulalem A., Shukor A., Siti H., Taiseer A., Arafat A., Maged N., Tusneem E., Hashim E., and Abdu S. (2022). Financial Fraud Detection Based on Machine Learning: A Systematic Literature Review. *Applied Sciences*, 12(19), 9637; https://doi.org/10.3390/app12199637

Anouche, M., & Boumaaz, Y. (2019). 'Customs risk management in developing countries: Foresight approach using big data'. *International Journal of Innovation and Applied Studies* ISSN 2028-9324,26(1): 58–68

Ashtiani, M. N.; Raahemi, B. (2021). 'Intelligent Fraud Detection in Financial Statements Using Machine Learning and Data Mining: A Systematic Literature Review'. *IEEE Access* 2021, 10, 72504–72525.

Bate, A., & Guedikouma, D. (2023). Incidence de la ZLECAF sur les revenus fiscaux au Togo. *African Multidisciplinary Tax Journal*, 1-19. doi:https://doi.org/10.47348/AMTJ/V3/i1a1

Bezabeh, B. (2019). 'Application of data mining for customs risk channel assignment: the case of Ethiopian revenue and customs authority'. (*Thesis in partial fulfillment of the requirements for the degree of master of science in computer science*.)

Cantens, T. (2015). 'Analyse miroir et fraude douanière.' *Document de recherche de l'OMD no 35*.

Carrère, C., & Grigoriou, C. (2015). 'Can mirror data help to capture informal international trade?' *Fondation pour les études et recherches sur le développement international*.

Geourjon, A-M Laporte, B., & Montagnat-Rentier, M. G. (2023). *The Use of Mirror Data by Customs Administrations: From Principles to Practice*. International Monetary Fund.

Geourjon, A.-M., & Laporte, B. (2012). 'La gestion du risque en douane: premières leçons tirées de l'expérience de quelques pays d'Afrique de l'Ouest'. *Revue d'économie du développement* 20(3): 67–82.

Grigoriou, C., Kalizinje, F., & Raballand, G. (2019). 'How helpful are mirror statistics for Customs reform? Lessons from a decade of operational use'. *World Customs Journal*, 13(2), 105–114.

Hala, Z., & Nojood, O. (2020). 'Fraud Detection in Credit Cards using Logistic Regression'. Department of Computer Science Tabuk University, Tabuk City Kingdom Saudi Arabia, International Journal of Advanced Computer Science and Applications, Vol, 11, No, 12.

Han, J., & Kamber, M. (2006). 'Data Mining: Concepts and Techniques'. 2nd ed, Morgan Kaufman publishers, San Francisco.

Khaled G., Pritheega M, (2021). «Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019». *Computer Science Review*, Volume 40, May 2021, 100402.

Mamo, D. (2013). 'Application of data mining technology to support fraud protection: the case of ethiopian revenue and custom authority'.

Ndikumana, L., & Boyce, J. (2021). 'Capital flight from Africa 1970-2018 New Estimates with Updated Trade Misinvoicing Methodology'. Political Economy Research Institute (PERI) University of Massachusetts-Amherst.

Qinghua Zheng, Yiming Xu, Huixiang Liu, Bin Shi, Jiaxiang Wang, Bo Dong (2024). 'A Survey of Tax Risk Detection Using Data Mining Techniques'. Journal homepage: www.elsevier.com/locate/eng

Sathya, R., & Abraham, A. (2013). 'Comparison of supervised and unsupervised learning algorithms for pattern classification'. *International Journal of Advanced Research in Artificial Intelligence*, 2(2), 34–38.

Shivakumar, S., & Sanjeev, C. (2014). 'Fraud Detection using Data Mining Techniques'. *International Journal of Innovations in Engineering and Technology*.

Tchila, P. (2020). «Analyse et gestion des risques douaniers de première ligne au Togo». Centre d'étude et de recherche sur le développement international (CERDI), Université Clermont Auvergne, 43p.

Uddin, S., Khan, A., Hossain, M., & Moni, M. (2019). 'Comparing different supervised machine learning algorithms for disease prediction'. *BMC medical informatics and decision making*, 19(1), 1–16.

Viet, C. (2015). "The Impact of Trade Facilitation on Poverty and Inequality: Evidence from low-and middle-income countries". *An International and Comparative Review, 24(3): 315-340.*

Walter D., Luca G., Giuseppe L., Lorenzo M., Fabrizio M., and Daniele P (2020). «Combining Network Visualization and Data Mining for Tax Risk Assessment». Received 27 December 2019, accepted 11 January 2020.

Zaki, C. (2011). 'Assessing the Global Effect of Trade Facilitation: Evidence from the Mirage Model'. Working Paper no, 659, Cairo: The Economic Research Forum.

https://doi.org/10.47348/AMTJ/V5/i1a1

ANNEXE

Code régime douanier	Nom régime douanier
1000	Exportation définitive
1093	Exportation ensuite de la zone franche
3070	Réexportation ensuite de l'entrepôt
4000	Mise à la consommation directe
4070	Mise à la consommation en suite de l'entrepôt
4077	Mise à la consommation en suite des Magasins et Aires de Dédouanement (MAD)
8000	Transit
8070	Transit ensuite de l'entrepôt
8077	Transit ensuit du MAD
9090F	Autres régimes
3080	Réexportation ensuite de transit
7070	Entrée en entrepôt ensuite d'un entrepôt
7080	Entrée en entrepôt ensuite du transit
9300	Entrée en zone franche

https://doi.org/10.47348/AMTJ/V5/i1a1