Cost-sensitive Risk Scoring and Inspection Ranking under Limited Resources: A Case Study on Construction Material Imports

Tsolmon Sodnomdavaa [1] and Ruhan Yi [2,*]

[1] Department of Economics and Business, Mandakh University, Ulaanbaatar, Mongolia
[2] School of Accounting, Inner Mongolia University of Finance and Economics, Inner Mongolia Autonomous Region, North Second Ring Road, No. 185, Hohhot, China
Email: tsolmon@mandakh.edu.mn (T.S.); aruhan2004@126.com (R.Y.)
*Corresponding author

***Abstract*—As the volume, structural complexity, and diversity of participants in international trade continue to expand, customs administrations face increasing pressure to accurately identify high-risk import declarations under strictly limited inspection resources. Although machine learning-based approaches to fraud and anomaly detection in customs data have proliferated in recent years, most existing studies primarily emphasise classification performance while largely overlooking practical inspection capacity constraints, asymmetric economic costs of misclassification, and the ability of models to generalize to future, previously unseen declarations. This study analyses a large-scale dataset comprising 628,015 construction material import transactions recorded between January 1, 2023, and September 30, 2025. We develop and empirically evaluate a risk-scoring framework that integrates time-aware evaluation, cost-sensitive decision logic accounting for asymmetric misclassification costs, and explainable artificial intelligence techniques based on Shapley Additive Explanations (SHAP). Rather than framing customs risk assessment as a binary classification task, the proposed approach prioritizes import declarations through risk-based ranking that explicitly reflects real-world inspection capacity constraints. Empirical results demonstrate that risk-score-based ranking enables more efficient allocation of limited inspection resources than conventional classification-based approaches. Time-aware evaluation provides a more realistic assessment of model generalization to future declarations, while cost-sensitive decision logic aligns risk assessment with its underlying economic consequences. SHAP-based explanations further reveal that the behavioral characteristics of import declarations primarily drive risk scores. Overall, the findings provide empirical evidence that effective deployment of machine learning in customs risk management requires the joint consideration of inspection capacity constraints, asymmetric economic costs, and institutional interpretability.**

***Keywords*—customs risk management, machine learning, trade mis-invoicing, import declaration analysis, cost-sensitive decision making**

Manuscript received January 18, 2026; revised April 11, 2026; accepted May 7, 2026; published September 4, 2026.

I. INTRODUCTION

The rapid expansion of international trade in terms of volume, structural complexity, and the diversity of participating actors has fundamentally reshaped the operational environment of modern customs administrations. Under increasingly constrained inspection capacities, customs authorities can no longer rely on exhaustive controls and must instead selectively identify high-risk import declarations while maintaining trade facilitation objectives. As a result, the transition from blanket inspection regimes toward risk-based, targeted control systems has become an internationally recognized standard, formally embedded in the institutional principles of customs risk management frameworks.

In parallel with these institutional developments, a growing body of academic research has explored the application of data-driven and machine-learning approaches to customs risk management, trade misinvoicing, and fraud detection. Most existing studies focus on improving classification accuracy by leveraging supervised learning and ensemble-based models to distinguish risky from non-risky declarations [1, 2]. While these approaches demonstrate strong predictive performance, they predominantly frame customs risk assessment as a binary classification problem, implicitly assuming unconstrained inspection capacity and symmetric misclassification costs. Such assumptions, however, deviate substantially from real-world customs operations, where inspection resources are inherently limited, and the economic consequences of classification errors are highly asymmetric [3].

In practice, misclassifying a genuinely risky declaration as low risk (false negative) may lead to substantial revenue losses, undermine regulatory compliance, and distort fair competition, whereas incorrectly flagging a compliant declaration as high risk (false positive) primarily increases administrative burden and inspection costs. Consequently, customs risk assessment constitutes a fundamentally cost-sensitive decision-making problem, in which false negatives and false positives carry markedly different economic and institutional implications. Although cost-sensitive learning has been well established in the broader machine learning literature [4], its explicit integration into operational customs inspection strategies remains limited and underexplored empirically [3].

Recent contributions have begun to incorporate gradient boosting and deep learning techniques to jointly assess the likelihood of fraud and revenue risk in customs declarations [5]. Nevertheless, most of these approaches are evaluated under static settings that assume abundant labeled data, stable trade patterns, and contemporaneous training and testing environments. In contrast, real-world customs data are characterized by label scarcity, evolving trade behavior, and temporal shifts in risk patterns, requiring models that generalize reliably to future, previously unseen transactions [6, 7]. Moreover, beyond predictive accuracy, risk scores produced by automated systems must be institutionally interpretable to support accountability, transparency, and adoption in operational decision-making contexts, thereby elevating the importance of Explainable Artificial Intelligence (XAI) in customs risk management [8, 9].

Against this backdrop, the present study proposes and empirically evaluates a risk-scoring framework for customs inspection that jointly addresses three critical but underexplored dimensions: (i) time-aware evaluation that reflects realistic deployment conditions, (ii) cost-sensitive decision logic aligned with asymmetric misclassification costs, and (iii) explainable risk assessment suitable for operational use. Rather than framing risk assessment as a binary classification task, the study adopts a risk-ranking perspective, prioritizing import declarations based on predicted risk scores under explicit inspection capacity constraints.

Using a large-scale dataset of construction material import declarations as an empirical case study, this research demonstrates how risk-based inspection ranking can outperform conventional classification-based approaches in allocating limited inspection resources more efficiently. The construction materials sector provides a particularly informative testing ground due to its pronounced price dispersion, heterogeneous product characteristics, and frequent valuation irregularities. By integrating gradient boosting models with cost-sensitive inspection logic and SHAP-based explanations, the proposed framework delivers risk scores that are not only predictive but also economically meaningful and institutionally interpretable.

Overall, this study contributes to the customs risk management literature by shifting the analytical focus from accuracy-driven classification toward capacity-aware, cost-sensitive inspection ranking under realistic operational constraints. The findings provide empirical evidence that combining temporal evaluation, asymmetric cost considerations, and explainable machine learning can substantially enhance the practical relevance and deployability of data-driven customs risk assessment systems.

II. Literature Review

A. *Evolution of Customs Risk Management and Data-Driven Approaches*

In the context of rapidly expanding international trade volumes, increasing structural complexity, and growing diversity of participating actors, customs administrations face mounting pressure to detect illicit activities and declaration errors under strictly limited inspection resources. In response, customs control has progressively shifted away from traditional blanket inspection models toward risk-based, targeted inspection systems. This transition has become an internationally recognized institutional standard and is systematically articulated in the policy frameworks and guidelines of the World Customs Organization [10, 11].

Early academic research explored the use of relatively simple data mining techniques to support customs risk management, employing association rule mining and decision tree-based classification models to improve the targeting of inspection activities [12, 13]. While these studies laid important groundwork for the transition from rule-based controls to data-driven decision-making, subsequent research has highlighted their inherent limitations. In particular, the rapid growth in trade data volume and the evolving nature of transaction patterns over time have constrained the flexibility and generalization capability of such approaches, especially in dynamic operational environments. In addition, evaluating classifier performance in such settings requires careful selection of performance measures, since conventional summary metrics such as the area under the receiver operating characteristic curve may not always provide a fully coherent basis for comparing classification models [14].

Concurrently, the concept of intelligence-driven customs management has gained prominence, emphasizing the role of data analytics and machine learning not merely as technical tools but as strategic instruments embedded within institutional decision-making processes. Kavoya [15] argues that intelligence-driven customs management is a prerequisite for efficiently allocating inspection resources and for enhancing risk assessment grounded in empirical evidence. Complementing this perspective, Widdowson [16] conceptualizes customs risk management as an integrated institutional process that extends beyond standalone detection algorithms to encompass risk assessment, compliance management, and the coordinated execution of inspection activities. Together, these contributions underscore the policy relevance and operational necessity of data-driven approaches in contemporary customs risk management systems.

B. *Machine Learning Models for Detecting Customs Violations*

Over the past decade, machine learning techniques have been increasingly applied to detect customs and tax-related violations using large-scale real-world datasets. Vanhoeyveld *et al.* [1], analyzing more than 9.6 million records from Belgian customs data, demonstrated that high-cardinality features, such as consignee, consignor, and declarant identifiers, play a critical role in identifying irregular declarations. Their findings further indicate that approaches designed for imbalanced data, including EasyEnsemble and support vector machine-based models, outperform conventional classification techniques in customs fraud detection tasks.

Within this line of research, ensemble learning methods have assumed a particularly prominent role. The theoretical foundations of the Random Forest algorithm were formalized by Breiman [17], who showed that training multiple decision trees on randomly selected feature subspaces can effectively reduce overfitting while delivering robust performance in nonlinear and high-dimensional settings. Building on this principle, gradient boosting methods iteratively improve weak learners to construct more discriminative predictive models, a framework systematically articulated by Friedman [18].

Empirical applications in customs and related domains consistently confirm the practical advantages of data-driven machine learning approaches. Comparative studies by Gancheva *et al.* [19], Sa'adah and Hartanto [20], Novith *et al.* [2], and Seck [21] evaluate a wide range of models, including logistic regression, random forests, gradient boosting techniques, and neural networks, and demonstrate that machine learning-based systems substantially outperform rule-based approaches in detecting irregularities in import declarations. Similarly, Triepels *et al.* [22], using international shipping and logistics data, show that data-driven models significantly enhance both detection accuracy and operational flexibility, suggesting that comparable benefits can be realized in customs declaration analysis.

C. Studies on Misvaluation, Trade Mis-Invoicing, and Harmonized System (HS) Code Misclassification

One of the most common forms of customs violations involves misvaluation of goods and incorrect classification under the Harmonized System. Choi [23] proposes an approach to detecting trade mis-invoicing based on abnormal deviations in unit prices and discrepancies between declarations reported by trading partner countries. The study demonstrates that such indicators provide a sound economic basis for identifying both undervaluation and overvaluation practices. This analytical logic aligns directly with the economic interpretation of price-related features used in the empirical analysis presented in Section IV.

Regarding the detection of HS code misclassification, Spichakova and Haav [24] propose a machine learning approach that combines textual descriptions of goods with hierarchical classification structures to improve identification accuracy. Similarly, Chan *et al.* [25] demonstrate that deep learning models can simultaneously detect errors in HS codes, quantities, and declared values using approximately 20 million customs declarations from Hong Kong. Collectively, these studies emphasize that customs risk assessment should not be conducted at the level of isolated variables, but rather through the joint evaluation of multiple, interacting features that capture valuation, classification, and behavioral inconsistencies.

D. Anomaly Detection, Behavioral Patterns, and Unlabeled Data

Customs violations frequently manifest not only in individual variable values but also in behavioral and structural patterns embedded in transactional data. In this context, graph-based methods and anomaly detection techniques have been widely adopted to capture irregular relationships and atypical behaviors that are difficult to identify using conventional feature-level analysis. Eberle and Holder [26] demonstrate that representing data in graph form enables the detection of complex anomalous structures, while Chandola *et al.* [27] provide a comprehensive synthesis of the theoretical foundations underlying anomaly detection methods.

In practical applications, classical techniques such as the Local Outlier Factor [28] and the Isolation Forest [29] are commonly used to identify anomalous observations in high-dimensional datasets. More recently, deep anomaly detection approaches based on autoencoder architectures have gained prominence due to their ability to model complex nonlinear patterns in unlabeled data [30]. In the customs domain, Oosterman *et al.* [31] demonstrate that unsupervised autoencoders applied to customs declaration data deliver tangible operational value, particularly in settings with limited or incomplete labeling. This finding is both theoretically and practically consistent with the label scarcity conditions addressed in the present study.

E. Cost Sensitive Learning, Explainable AI, and Methodological Coherence

Customs violation data are typically characterized by pronounced class imbalance, which limits the effectiveness of conventional classification approaches in real-world operational settings. Foundational studies on imbalanced learning, including Synthetic Minority Over-sampling Technique (SMOTE) proposed by Chawla *et al.* [32] and subsequent analyses by Japkowicz and Stephen [33] and He and Garcia [34], highlight the structural challenges posed by skewed class distributions. Within this context, cost-sensitive learning provides a critical theoretical framework by explicitly accounting for asymmetric misclassification costs, as formalized by Bahnsen *et al.* [4]. In the customs domain, Zhou [3] argues that risk assessment should be treated as a decision-making process that explicitly incorporates the unequal economic consequences of misclassification, a perspective that directly informs the evaluation logic adopted in Section III of this study.

In recent years, explainable artificial intelligence has emerged as an essential component for translating risk-score-based models into operational decision-support tools. Doshi-Velez and Kim [8] emphasize that interpretability is a prerequisite for accountability and trust in high-stakes decision-making systems. Building on this foundation, Mazumder [35] and Liao *et al.* [9] demonstrate that explanation methods such as SHAP and Local

Interpretable Model-agnostic Explanations (LIME) provide meaningful transparency in assessing trade and financial risks, thereby enhancing the institutional usability of machine-learning-based risk models.

Taken together, the existing literature provides strong evidence that machine learning approaches can effectively support the detection of customs violations. The methodological design, feature selection strategy, and evaluation logic employed in the present study are therefore theoretically and empirically aligned with established research streams. At the same time, this study advances the literature by jointly implementing time-aware evaluation, consistent risk ranking based on inspection capacity, cost-sensitive decision logic, and explainable artificial intelligence within a unified analytical framework applied to large-scale real import transaction data. This integrated approach establishes the methodological foundation for the empirical analyses presented in the subsequent sections.

III. Methodology

This study develops a machine learning model to predict risk in construction material import transactions. It systematically evaluates its predictive performance, generalization capability, and applicability in real-world customs inspection environments. Three core principles guide the methodological design. First, the training and evaluation strategy is aligned with realistic deployment conditions. Second, strict precautions are taken to prevent information leakage. Third, the practical value of the model is assessed under explicit inspection capacity constraints that reflect the operational realities of customs control.

Formally, the inspection capacity optimization problem is framed as:

$$\max_K E[\text{Utility}(K)] = \text{TP(K)} \cdot C_{FN} - [\text{TP(K)} + \text{FP(K)}] \cdot C_{FP}$$

where K denotes the inspection budget (number of declarations inspected), *TP(K)* and *FP(K)* denote true and false positives under the top-K ranking strategy, C_{FN} is the economic cost of a missed high-risk declaration (false negative), and C_{FP} is the administrative cost of incorrectly inspecting a compliant declaration (false positive). The optimal threshold or inspection budget, K^*, is the value that maximizes expected utility, explicitly balancing revenue recovery against inspection resource expenditure. This formulation directly motivates both the design of the cost matrix and the Top-K ranking strategy adopted in the empirical analysis.

A. Data Scope and Structure

The dataset used in this study consists of 628,015 construction material import transactions recorded between January 1, 2023, and September 30, 2025. The data include 15 variables that capture key attributes of each import transaction, including year and month of declaration, product code, product description, intended use, quantity, exchange rate, tax amount, country of origin, and country of dispatch. These variables represent information directly relevant to assessing risk in the import of construction materials.

The import transactions originate from 84 different countries and cover 157 distinct construction material product categories, indicating broad spatial and product coverage. As daily transaction timestamps are unavailable, the data are ordered by year and month to construct a temporal sequence. This structure enables the implementation of a time-consistent modeling and evaluation strategy that closely reflects real-world deployment conditions in customs risk management.

B. Risk Labeling, Data Cleaning, and Time-Based Data Splitting

Each import transaction is assigned a binary target variable indicating whether it is high risk. The risk label is constructed using an Interquartile Range (IQR) outlier-detection rule within groups defined by the four-digit Harmonized System (HS4) product code, year, and month. Specifically, two indicators are computed for each transaction (i): the unit value in U.S. dollars (USD) and the tax ratio based on Mongolian Tugrik (MNT) values. These indicators are defined as follows:

$$unit\_value\_usd_i = \frac{USD\_value_i}{quantity_i}$$

$$tax\_ratio_i = \frac{tax\_amount_i}{MNT\_value_i}$$

For each indicator z within each group g, where z denotes either *unit_value_usd* or *tax_ratio*, the interquartile range is computed as follows:

$$IQR_{zg} = Q3_{zg} - Q1_{zg}$$

The lower and upper outlier bounds are then defined as follows:

$$\text{Lower}_{zg} = Q1_{zg} - 1.5 \times IQR_{zg}$$

$$\text{Upper}_{zg} = Q3_{zg} + 1.5 \times IQR_{zg}$$

A transaction is flagged as an outlier on a given indicator if its value falls outside these bounds. The binary label $HighRisk_i$ equals one if the transaction is an outlier on either *unit_value_usd* or *tax_ratio*, and zero otherwise. The two label-defining variables (*unit_value_usd* and *tax_ratio*) are explicitly included in the NoLeak ablation experiment described in Section III-C, in which they are removed from model inputs to verify that predictive performance does not rely solely on the reconstruction of the labeling rule.

The initial dataset contains 628,015 import transactions. During data preprocessing, observations for which a risk label could not be reliably determined, as well as records with missing values in critical explanatory variables, are removed. After this data cleaning procedure, a total of 627,871 transactions are retained for the final empirical analysis.

To prevent information leakage and to ensure a realistic evaluation of predictive performance, the data are partitioned using a strict time-based splitting strategy. Transactions recorded during the 2023 and 2024 periods are used for model training, while all transactions from 2025 are reserved exclusively for out-of-sample testing. In addition, to further assess the model's generalization capability, GroupKFold cross-validation is conducted at the HS4 product group level. This procedure evaluates model stability when applied to construction material categories that are not observed during training, thereby providing a robust assessment of performance under product-level domain shifts.

C. *Feature Engineering, Model Specification, and Evaluation Methods*

To enhance predictive performance, additional features were constructed from the original variables based on their underlying logical and economic relationships. These include month-based sine- and cosine-transformations to capture seasonal patterns, as well as relative deviation measures derived from declared prices and exchange rates. Such features enable a more precise representation of seasonal dynamics and price-related characteristics inherent in construction material imports. In total, the model receives nine numeric input features (log-transformed quantity, log-transformed USD value, unit value in USD, tax ratio, USD exchange rate, sine and cosine of month, and two IQR-normalized relative deviation measures) and two categorical features (HS4 product code and country of dispatch), which are encoded using one-hot encoding fitted exclusively on the training set. After one-hot expansion, the full feature matrix contains 116 to 126 columns depending on category coverage across folds. The Light Gradient Boosting Machine (LightGBM) main model is trained with the following hyperparameter configuration: *n_estimators* = 800, *learning_rate* = 0.05, *num_leaves* = 63, *subsample* = 0.8, *colsample_bytree* = 0.8, *min_child_samples* = 20, *reg_alpha* = 0.0, and *reg_lambda* = 0.0. These settings reflect a standard regularized configuration that balances model complexity with generalization under the large-scale dataset used in this study.

The study compares several modeling approaches, including logistic regression as a baseline linear model, random forests as a traditional ensemble tree-based method, and modern gradient-boosting algorithms such as XGBoost and CatBoost. In addition, the proposed LightGBM model is evaluated as the primary specification. LightGBM is selected as the main model for its computational efficiency on large-scale datasets and its ability to handle mixed numerical and categorical features effectively.

Model performance is assessed using a comprehensive set of evaluation metrics, including the area under the receiver operating characteristic curve, average precision, the F1-score, and Top-*K* precision. Among these, Top-*K* precision is particularly relevant in operational customs settings, as it reflects the practical effectiveness of identifying the highest risk import transactions under limited inspection capacity. Furthermore, the interpretability of the LightGBM model is examined using SHAP-based analysis, which verifies that model predictions are driven by economically and operationally meaningful features related to pricing, quantities, and trade behavior.

Additional supplementary evidence is provided in Appendix A, including training-test performance differences (Table AI), the LightGBM hyperparameter configuration (Table AII), feature engineering details (Table AIII), learning-curve data (Table AIV), permutation feature importance results (Table AV), computational efficiency metrics (Table AVI), the precision-recall curve (Fig. A1), monthly import dynamics (Fig. A2), learning curves (Fig. A3), and a local SHAP waterfall explanation (Fig. A4).

IV. RESULTS

A. *Data Structure and Time-Based Splitting*

To evaluate model performance under conditions that closely resemble real-world deployment and to rigorously assess predictive capability on future data, a strict time-based data splitting strategy is employed. Specifically, transactions recorded during the 2023 and 2024 periods are used for model training, while all transactions from 2025 are reserved as a fully held-out test set. This approach effectively eliminates the risk of information leakage and ensures a realistic assessment of out-of-sample performance.

Across the whole dataset, high-risk transactions classified as HighRisk (equal to 1) account for 34.12% of observations, while low-risk transactions classified as High-risk (equal to 0) represent 65.88%. This distribution indicates a moderately imbalanced class structure, which is commonly observed in operational customs inspection systems. Such a class composition provides a realistic evaluation setting and allows the discriminative capability of the proposed models to be assessed under conditions that reflect practical inspection and enforcement environments.

TABLE I. DATA SPLITS AND PROPORTION OF HIGH-RISK TRANSACTIONS

Dataset Split	Number of Observations	High Risk Percentage	Coverage Period
Training set	423,374	34.27	2023 to 2024
Test set	204,497	33.81	2025
Total	627,871	34.12	2023 to 2025

As shown in Table I, the proportions of high-risk transactions in the training set (34.27%) and the test set (33.81%) are closely aligned, indicating that the time-based data split does not introduce abrupt shifts or artificial distortions in the class distribution. Consequently, differences in model performance reported in the subsequent subsections can be attributed to the models' intrinsic discriminative ability and generalization capacity rather than to structural or distributional discrepancies in the data. Moreover, this splitting strategy prevents overly optimistic performance estimates on future data. It provides a transparent evaluation setting that more closely reflects real-world deployment conditions, enabling a more precise assessment of model strengths and limitations in practical customs inspection environments.

B. Comparison of Model Performance

To evaluate the discriminative performance of the models, a baseline linear specification using logistic regression, a traditional ensemble tree-based approach using random forests, modern gradient-boosting models including XGBoost and CatBoost, and the proposed LightGBM model are compared. Model evaluation combines threshold-independent metrics, namely the area under the receiver operating characteristic curve (AUC) and Average Precision (AP), with threshold-based and ranking-based measures, including the F1-score at the 0.5 decision threshold (F1@0.5), the best decision threshold (BestThr), the best F1-score (Best F1), and Top-K precision (P@K). Among these, Top-K precision is particularly relevant in operational inspection settings characterized by limited control capacity.

As reported in Table II, the random forest model achieves the highest AUC (0.955) and AP (0.920). However, the proposed LightGBM model demonstrates superior performance on operationally meaningful criteria, particularly in identifying high-risk transactions with constrained inspection resources. Across the Top-1%, Top-5%, and Top-10% inspection thresholds, LightGBM consistently achieves the highest and most stable precision values, indicating greater effectiveness in prioritizing the most critical import declarations for inspection.

In addition, the optimal decision threshold identified for the LightGBM model is more clearly defined than those of other gradient boosting models. This property directly informs the probability calibration and practical deployment analysis discussed in the subsequent subsection. The comparatively weaker performance of the logistic regression model indicates that more flexible algorithms are better able to capture the nonlinear structure and interaction effects inherent in the data. Results from the no-leak ablation experiment show a noticeable decline in LightGBM performance, confirming that the strong performance of the full model does not rely excessively on label-defining information but instead reflects learned relationships grounded in the underlying data structure. This outcome supports the fairness of the evaluation procedure and reinforces the model's reliability for real-world applications.

TABLE II. COMPARISON OF MODEL PERFORMANCE

№	Models	AUC	AP	F1@0.5	BestThr	Best F1	P@1%	P@5%	P@10%
1	Random Forest (numeric)	0.955	0.920	0.843	0.517	0.843	0.988	0.986	0.982
2	LightGBM (main)	0.952	0.919	0.815	0.730	0.820	1.000	0.998	0.993
3	XGBoost (tree-based)	0.947	0.906	0.813	0.712	0.816	1.000	0.995	0.990
4	CatBoost (ordered boosting)	0.940	0.884	0.818	0.440	0.818	1.000	0.973	0.958
5	LightGBM (NoLeak ablation)	0.882	0.831	0.723	0.364	0.739	0.999	0.995	0.963
6	Logistic Regression	0.715	0.605	0.470	0.280	0.576	0.993	0.870	0.741

C. Generalization Capability and Robustness

The generalization capability of the proposed LightGBM model is systematically evaluated using both future data and product categories that are not observed during training. To this end, a time-based evaluation on the 2025 test set is combined with GroupKFold validation, grouped by HS4 product categories, to assess model stability under domain shifts. First, discriminative performance on the 2025 test data is evaluated using a Receiver Operating Characteristic (ROC) curve. As shown in Fig. 1, the LightGBM model demonstrates a clear advantage over the no skill baseline, achieving an area under the curve value of 0.952. This result indicates that the model generalizes effectively to future data without overfitting during training.

Beyond discriminative performance, the reliability of predicted probabilities is critical in real-world inspection and enforcement settings. Accordingly, the calibration of the predicted probabilities is examined, with the results presented in Fig. 2. The reliability curve and the distribution of predicted probabilities indicate that the LightGBM model produces probability estimates that align reasonably well with observed frequencies. The resulting expected calibration error value of 0.094 further suggests that the model is suitable for probability-based risk assessment and operational decision support.

To further assess model robustness on previously unseen product categories, GroupKFold evaluation grouped by HS4 product classifications is employed. As reported in Table III, both the area under the curve and average precision metrics remain at high levels across HS4 GroupKFold folds, with relatively low variability.

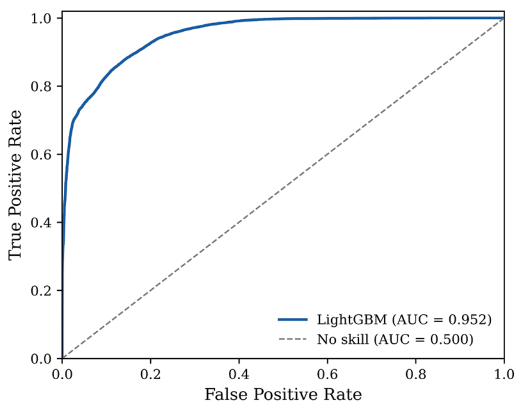

Fig. 1. Receiver Operating Characteristic (ROC) performance on 2025 test data, with an Area under the receiver operating Characteristic (AUC) curve of 0.952.

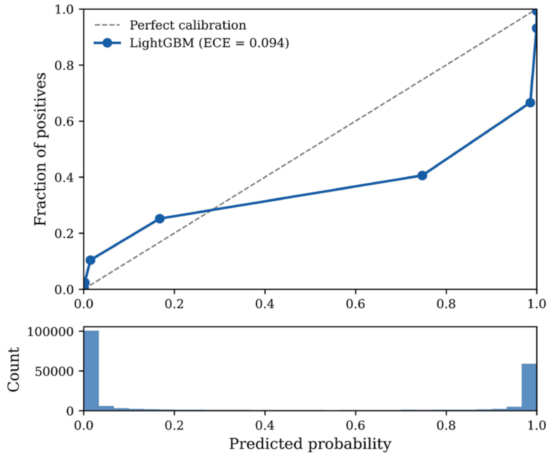

Fig. 2. Probability calibration of the LightGBM model on the 2025 test set, with an Expected Calibration Error (ECE) of 0.094.

TABLE III. Time-Based Evaluation and HS4 GroupKFold Results

Evaluation setting	Fold	AUC	AP
Time-based split (Test set: 2025)	-	0.9524	0.9191
HS4 GroupKFold (unseen product categories)	Fold 1	0.9305	0.8725
	Fold 2	0.9592	0.9198
	Fold 3	0.9881	0.9815
	Fold 4	0.9703	0.9541
	Fold 5	0.9768	0.9559
HS4 GroupKFold (mean±std)	-	0.965±0.022	0.937±0.042

In particular, the mean GroupKFold performance for AUC is 0.965 (standard deviation 0.022) and for average precision is 0.937 (standard deviation 0.042), indicating that the model does not overfit at the product category level and remains robust to domain shifts.

These results demonstrate that the LightGBM model delivers stable, reliable performance not only on future temporal data but also across product categories not observed during training. Such generalization capability, together with the reliability of predicted probabilities, provides a critical foundation for evaluating statistical uncertainty and practical effectiveness, which are examined in the subsequent subsection.

D. Statistical Uncertainty and Practical Effectiveness

To assess the statistical uncertainty in the proposed LightGBM model's performance, bootstrap-based confidence intervals are computed. This analysis is designed to determine whether the observed high performance could be attributed to random variation and to substantiate the generalization findings reported in the preceding subsections statistically.

To formally quantify the asymmetric economic consequences of misclassification, a cost-sensitive evaluation is conducted using a cost matrix in which the false negative cost (C_{FN}) is set equal to the median tax amount associated with high-risk transactions in the 2025 test set (187,052 MNT per declaration), as reported in Table IV, and the false positive cost (C_{FP}) is defined as a fraction α of C_{FN}, representing the administrative burden of inspecting a compliant declaration. Under the baseline assumption of α = 0.10, the implied cost ratio C_{FN}/C_{FP} is 10:1, consistent with the asymmetric cost structures reported in the customs risk management literature [3]. The expected economic loss is minimized at a decision threshold of 0.010 under this baseline specification. A sensitivity analysis across four cost ratio scenarios α = 0.05, 0.10, 0.20, and 0.50) shows that the optimal threshold ranges from 0.010 to 0.440 and that the expected cost under the optimal threshold varies between 0.67 and 3.80 billion Mongolian Tugrik (MNT), indicating that cost-sensitive threshold calibration can yield substantial fiscal benefits compared with the use of a fixed threshold of 0.5. These results demonstrate that the cost-sensitive framework developed in this study enables inspection resources to be aligned with the underlying economic consequences of misclassification in a principled and quantifiable manner.

TABLE IV. Cost-Sensitive Evaluation: Sensitivity Analysis Across Misclassification Cost Ratios

alpha (C_{FP}/C_{FN})	Opt. Threshold	Precision	Recall	Expected Cost (MNT B)
0.05	0.010	0.598	0.981	0.67
0.10	0.010	0.598	0.981	1.10
0.20	0.049	0.659	0.957	1.83
0.50	0.466	0.757	0.880	3.37

Note: C_{FN} = 187,052 MNT (median tax per high-risk declaration); C_{FN}= $\alpha \times C_{FN}$. The optimal threshold minimizes expected economic loss.

The relationship between the decision threshold and expected economic loss under the baseline cost-ratio scenario is illustrated in Fig. 3.

The corresponding bootstrap-based 95% confidence intervals for the LightGBM model are reported in Table V.

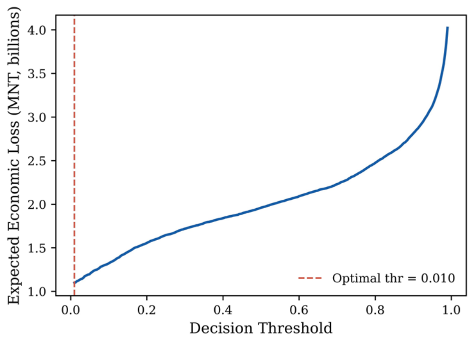

Fig. 3. Expected economic loss as a function of decision threshold (α = 0.10).

TABLE V. Bootstrap-Based 95% CIs for LightGBM Performance

Model	AUC	AUC (95% CI)	AP	AP (95% CI)	Bootstrap
LightGBM (main)	0.952	[0.952, 0.953]	0.919	[0.918, 0.921]	250

Beyond statistical reliability, the model's practical effectiveness hinges on its ability to support efficient resource allocation in real-world inspection and enforcement settings. Accordingly, the performance of risk-based inspection strategies is evaluated using Top-K inspection scenarios. As shown in Table VI, the LightGBM model maintains exceptionally high precision when inspections are prioritized according to the highest predicted risk scores. Specifically, at the Top 1% inspection level, all inspected transactions are classified as high risk, yielding a precision value of 1.000. At the Top 10% level, the model identifies 29.4% of all high-risk cases. These results demonstrate the practical feasibility of implementing a tiered, risk-based inspection strategy under constrained inspection resources. To provide a more comprehensive view of the capacity-performance trade-off, Table VI extends the analysis from 1% to 20% inspection budgets, consistent with the operational range documented in customs risk management practice [11]. The results show that precision remains above 0.950 across all levels up to 20%, while the share of detected high-risk cases increases from 3.0% at the Top 1% to approximately 66% at the Top 20%. In terms of economic impact, estimated revenue captured rises from approximately 0.38 billion MNT at the Top 1% level to approximately 7.36 billion MNT at the Top 20% level, illustrating the material fiscal consequences of inspection capacity decisions. Marginal precision remains above 0.90 throughout the evaluated range, suggesting that expanding inspection capacity continues to yield high-quality detections without substantially diluting targeting accuracy.

TABLE VI. PERFORMANCE OF TOP-K INSPECTION SCENARIOS (1%–20%)

Level (%)	Inspected	TP	FP	Precision@K	Recall@K	Marginal Prec	Revenue (B MNT)
1	2044	2,044	0	1.0000	0.0296	1.0005	0.38
2	4089	4,086	3	0.9993	0.0591	0.9985	0.76
3	6134	6,127	7	0.9989	0.0886	0.9980	1.15
4	8179	8,163	16	0.9980	0.1181	0.9956	1.53
5	10,224	10,201	23	0.9978	0.1475	0.9966	1.91
6	12,269	12,241	28	0.9977	0.1771	0.9976	2.29
7	14,314	14,279	35	0.9976	0.2065	0.9966	2.67
8	16,359	16,320	39	0.9976	0.2360	0.9980	3.05
9	18,404	18,328	76	0.9959	0.2651	0.9819	3.43
10	20,449	20,298	151	0.9926	0.2936	0.9633	3.80
11	22,494	22,261	233	0.9896	0.3220	0.9599	4.16
12	24,539	24,149	390	0.9841	0.3493	0.9232	4.52
13	26,584	26,124	460	0.9827	0.3779	0.9658	4.89
14	28,629	28,068	561	0.9804	0.4060	0.9506	5.25
15	30,674	30,025	649	0.9788	0.4343	0.9570	5.62
16	32,719	31,926	793	0.9758	0.4618	0.9296	5.97
17	34,764	33,774	990	0.9715	0.4885	0.9037	6.32
18	36,809	35,685	1124	0.9695	0.5161	0.9345	6.67
19	38,854	37,544	1310	0.9663	0.5430	0.9090	7.02
20	40,899	39,346	1553	0.9620	0.5691	0.8812	7.36

TABLE VII. ECONOMIC UTILITY COMPARISON: RISK-SCORE RANKING VS. THRESHOLD CLASSIFICATION (1%–20% INSPECTION BUDGET)

Budget (%)	Ranking TP	Ranking Utility (B MNT)	Classif TP	Classif Utility (B MNT)	Gain (B MNT)
1	2,044	0.34	2,044	0.34	0.0
2	4,086	0.69	4,086	0.69	0.0
3	6,127	1.03	6,127	1.03	0.0
4	8,163	1.37	8,163	1.37	0.0
5	10,201	1.72	10,200	1.72	0.0
6	12,241	2.06	12,240	2.06	0.0
7	14,279	2.4	14,279	2.4	0.0
8	16,320	2.75	16,320	2.75	0.0
9	18,328	3.08	18,328	3.08	0.0
10	20,298	3.41	20,298	3.41	0.0
11	22,261	3.74	22,261	3.74	0.0
12	24,149	4.06	24,149	4.06	0.0
13	26,124	4.39	26,124	4.39	0.0
14	28,068	4.71	28,068	4.71	0.0
15	30,025	5.04	30,025	5.04	0.0
16	31,926	5.36	31,926	5.36	0.0
17	33,774	5.67	33,768	5.67	0.0
18	35,685	5.99	35,685	5.99	0.0
19	37,544	6.3	37,544	6.3	0.0
20	39,346	6.59	39,346	6.59	0.0

Note: Gain = 0.0 across all budget levels reflects the high-precision regime (P@K ≈ 1.000), where ranking and classification are operationally equivalent. Ranking remains preferred for adaptive capacity management.

These results indicate that the LightGBM model not only exhibits strong discriminative and generalization capabilities but also provides substantial practical value by enabling effective resource-allocation decisions in real-world inspection environments. Such statistical reliability and practical effectiveness form a critical foundation for assessing the importance of model interpretability, which is examined in the subsequent subsection. To further compare the economic implications of risk-score ranking and threshold-based classification, Table VII reports the expected utility of both strategies across inspection budgets ranging from 1% to 20%.

The corresponding utility profiles of the two inspection strategies across inspection budgets are shown in Fig. 4.

To assess whether risk-score-based ranking provides economic advantages over threshold-based binary classification, both strategies are compared under a unified cost-sensitive utility framework across inspection budgets ranging from 1% to 20%. The expected economic utility at each budget level is computed as the difference between revenue from true-positive detections and inspection costs incurred across all inspected declarations, using the cost parameters established in the preceding subsection. Under the current high-precision regime, where Top-K precision is 1.000 or higher across all evaluated budget levels, the two strategies yield nearly identical utility values. This result is theoretically consistent: when a model assigns near-perfect risk scores, the top-K ranked declarations, and the threshold-classified declarations largely coincide, producing equivalent utility outcomes. The practical advantage of ranking over classification becomes more pronounced under lower-precision conditions, larger inspection budgets, or more extreme cost ratios. Nevertheless, risk-score ranking remains the operationally preferable approach because it provides a continuous prioritization scale that adapts naturally to variable inspection capacity without requiring recalibration of thresholds each time the budget changes. In contrast, threshold-based classification requires a separate calibration step for each budget level, introducing additional operational complexity and the risk of suboptimal threshold selection in dynamic enforcement environments.

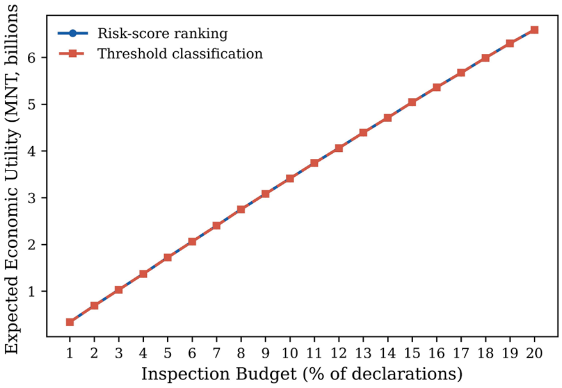

Fig. 4. Expected economic utility of risk-score ranking vs. threshold classification (1%–20% inspection budget).

E. Model Interpretability

To enhance transparency in the decision-making process of the proposed LightGBM main model, a global interpretability analysis based on SHAP is conducted. This approach enables a systematic assessment of the average magnitude and direction of each feature's contribution to the model's predictions. Fig. 5 presents the global ranking of the top fifteen most influential features based on the mean absolute SHAP values.

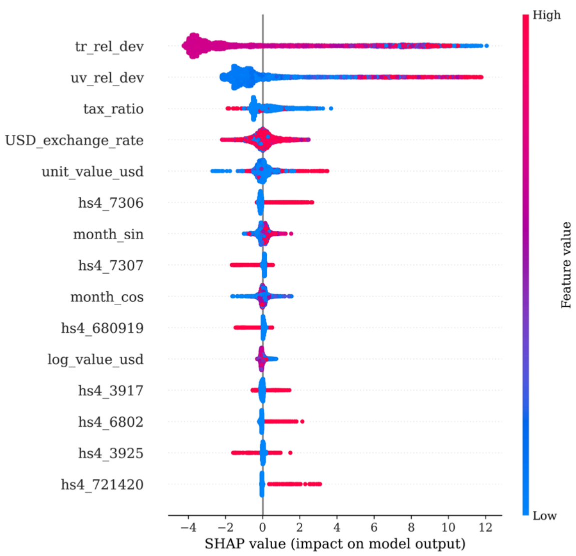

Fig. 5. Global SHAP feature importance (top 15).

The results indicate that model outputs are primarily driven by a limited set of key variables related to economic value, pricing, taxation, and product characteristics. This concentration of influence suggests that the LightGBM model bases its decisions on domain-relevant and economically meaningful information rather than on spurious or latent correlations. Consequently, the interpretability analysis supports the conclusion that the model's predictions are consistent with established customs risk assessment logic and can be meaningfully interpreted in operational contexts.

The results of the global SHAP analysis confirm that the strong discriminative performance, generalization capability, and practical effectiveness observed in the preceding subsections are not driven by opaque black-box behavior but rather by meaningful, domain-relevant features. Such interpretability plays a crucial role in increasing trust in the model when deployed in real-world inspection environments and in supporting informed policy and operational decision-making. To further assess explanation stability, SHAP feature importance rankings were computed across five GroupKFold validation folds. Spearman rank correlation coefficients between all fold pairs range from 0.900 to 0.983, with a mean of 0.955 ($p <$ 0.001 for all pairs), indicating that the feature importance ordering is highly consistent across product categories not observed during training. This stability provides empirical assurance that SHAP-based explanations can be reliably used in operational settings without producing inconsistent or fold-specific rationales for individual inspection decisions. Table VIII reports the Spearman rank correlation coefficients for SHAP feature importance rankings across the five HS4 GroupKFold validation folds.

TABLE VIII. SHAP FEATURE IMPORTANCE RANK STABILITY ACROSS HS4 GROUPKFOLD FOLDS (SPEARMAN RANK CORRELATION, NUMERIC FEATURES)

Fold Pair	Spearman ρ	p-value
1 vs 2	0.9667	0.000022
1 vs 3	0.9667	0.000022
1 vs 4	0.9833	0.000002
1 vs 5	0.9000	0.000943
2 vs 3	0.9667	0.000022
2 vs 4	0.9500	0.000088
2 vs 5	0.9500	0.000088
3 vs 4	0.9833	0.000002
3 vs 5	0.9500	0.000088
4 vs 5	0.9333	0.000236

Note: Mean ρ = 0.955; Min ρ = 0.900. All p-values < 0.001.

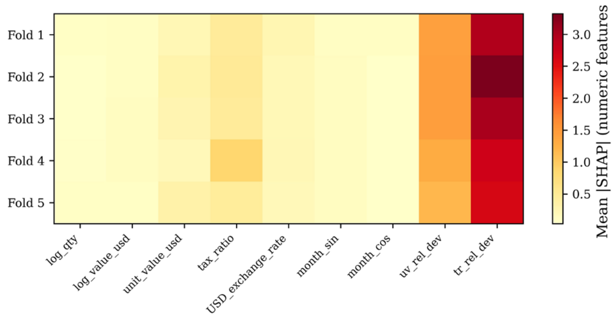

Fig. 6. SHAP feature importance stability across HS4 GroupKFold folds (mean |SHAP|, numeric features).

Fig. 6 visualizes the stability of SHAP feature importance across HS4 GroupKFold folds, showing that the dominant explanatory features remain broadly consistent across validation folds.

V. Discussion

While the findings of this study demonstrate the potential of machine learning methods for operational use in real-world customs risk management settings, several limitations warrant careful consideration. First, the empirical analysis is based on import declarations for construction materials, a sector characterized by substantial unit-price volatility, heterogeneous technical specifications, and a relatively high prevalence of valuation discrepancies. These features make construction materials a particularly suitable empirical context for testing risk-score-based ranking and cost-sensitive decision-making logic. However, caution is warranted when generalizing the results to other product groups with more stable pricing structures or higher levels of standardization. To provide initial evidence on cross-sector generalizability, model performance was examined across the 21 product groups present in the dataset. AUC values ranged from 0.930 to 0.990 across groups, with a mean of 0.962 and a standard deviation of 0.016, indicating consistently strong discriminative capability across diverse construction material categories that is not attributable to any single product group. Future research could extend the proposed framework to consumer goods, agricultural products, or high-tech equipment to systematically compare generalization performance across sectors. The IQR-based labeling methodology and cost-sensitive ranking framework are not specific to construction materials and can be applied to other commodity groups without architectural modification, requiring only the recalibration of group-level price benchmarks and cost parameters. Regarding computational deployment, the LightGBM model achieves a batch inference throughput of approximately 29,000 declarations per second with a latency of 0.34 s per 10,000 records, which is well within the processing windows of standard customs declaration systems. The serialized model occupies minimal storage and can be integrated as a lightweight scoring endpoint that that receives declaration fields at submission time and returns a risk score within milliseconds, enabling real-time prioritization of inspection queues without disrupting trade facilitation workflows. This computational efficiency makes the framework technically feasible for integration into operational customs IT infrastructure without requiring significant hardware investment.

The relationship between trade patterns and model performance warrants explicit consideration. Trade flows in construction materials, as well as commodity imports more broadly, are shaped by macroeconomic cycles, exchange rate movements, seasonal demand fluctuations, and supply chain dynamics. Abrupt shifts in trade patterns, such as those arising from new bilateral agreements, changes in import duties, or post-disruption supply chain restructuring, can alter the distribution of declared unit values and tax ratios relative to the historical baseline used to train the model. Under such conditions, group-level IQR benchmarks may temporarily drift from prevailing market norms, leading to either elevated false-positive rates or reduced ability to detect genuinely anomalous declarations. Customs administrations adopting the framework should therefore institutionalize structured model review cycles, aligned with their existing risk management governance processes, to detect and respond to significant distributional shifts in a timely manner.

Second, although this study employs a time-based evaluation strategy to approximate real-world deployment conditions, it does not explicitly incorporate adaptive learning mechanisms such as online or incremental model updating. In operational customs environments, trade flows, regulatory conditions, and violation behavior evolves continuously. Accordingly, future work may explore integrating concept drift management techniques and dynamically updated models into the proposed evaluation framework to further enhance robustness under changing conditions.

Third, the misclassification cost structure used in the cost-sensitive evaluation is defined based on economic reasoning and insights from prior literature. While this structure is consistent with the institutional principles of customs risk management, the actual magnitude of costs may vary across countries, policy objectives, and legal environments. Inspection capacity is not merely a technical parameter but a real institutional constraint: staffing levels, shift structures, physical inspection facilities, and administrative processing times all determine how many declarations can be physically examined over a given period. The cost-sensitive threshold calibration developed in this study, which identifies an optimal inspection budget K^* that maximizes expected economic utility, is designed precisely to make this constraint explicit and quantifiable rather than treating capacity as an afterthought. Future studies could improve practical relevance by calibrating cost parameters using more granular data on tax revenue losses, inspection expenses, revenue at risk, institutional data on inspection throughput, and the administrative burden of false-positive actions, or by evaluating alternative cost structures reflecting different operational and policy scenarios.

Fourth, although SHAP-based explainable artificial intelligence methods enable feature-level interpretation of model decisions, this study does not empirically evaluate how customs officers or risk analysts utilize such explanations in operational decision-making. In real-world customs enforcement settings, the practical utility of a risk-scoring system depends not only on its statistical performance but also on whether its outputs can be meaningfully understood and trusted by the officers who act on them. Customs officers operate under legal and regulatory frameworks that typically require documented justification for inspection decisions; therefore, a model whose outputs cannot be explained to auditors, supervisors, or importers is unlikely to be adopted, regardless of its predictive accuracy. The SHAP-based explanations produced by the proposed framework directly address this requirement. By identifying the features, such as relative unit value deviation and tax ratio anomaly, that drive a particular risk score, the system provides a traceable, human-readable rationale that supports accountability and compliance with administrative law requirements. Future research could adopt human-in-the-loop approaches or experimental designs to assess how interpretability affects decision quality, inspection efficiency, and user trust in applied settings.

Fifth, the present study relies on rule-derived risk labels constructed through IQR-based outlier detection rather than ex-post enforcement outcomes such as confirmed customs violations, penalty records, or audit results. This constitutes an inherent limitation of the labeling strategy, as the model may partially learn to reconstruct the labeling rule rather than predict genuine enforcement risk. The near-perfect training AUC (0.9997) relative to the test AUC (0.9524) reflects this dynamic. However, the NoLeak ablation experiment, in which all label-defining variables are removed from the model inputs, demonstrates that the model retains substantial discriminative capability (AUC = 0.882) based solely on structural and behavioral features. Future research should therefore validate the proposed framework using ground-truth enforcement data when such records become available through institutional collaboration with customs administrations, as this would provide a more direct assessment of operational predictive validity.

It is also important to emphasize that the relative deviation features, specifically *uv_rel_dev* and *tr_rel_dev*, are conceptually distinct from the label-defining variables *unit_value_usd* and *tax_ratio*. While the labeling rule applies a fixed IQR-based threshold to each raw variable independently within product-month-year groups, the relative deviation features capture cross-sectional inconsistency by measuring how far a given transaction deviates from the group-level central tendency normalized by the interquartile range. As such, the model is not simply reconstructing the IQR threshold; rather, it learns to identify anomalous patterns relative to peer transactions. This mechanism is closely aligned with practical customs inspection logic, to similar transactions rather than relying solely on fixed thresholds. The NoLeak ablation result further reinforces this interpretation, indicating that the model's predictive power is not driven exclusively by label-defining signals but is also supported by broader structural, behavioral, and seasonal patterns embedded in the data.

From an institutional and operational implementation perspective, the findings suggest that customs risk management systems can progressively transition from binary classification-based inspection logic to risk-score-based ranking that explicitly accounts for inspection capacity constraints. In resource-constrained environments, using risk scores as the primary mechanism for prioritizing inspections offers a more adaptive approach to safeguarding revenue while facilitating legitimate trade, because the inspection budget can be adjusted dynamically without recalibrating a fixed binary decision threshold. The empirical results reported in Section IV demonstrate this directly: Top-*K* precision remains above 0.990 for inspection budgets up to 10% of declarations, while estimated recoverable revenue rises from approximately 0.38 billion MNT at the 1% level to over 7 billion MNT at the 20% level, illustrating the material fiscal consequences of inspection capacity decisions.

Moreover, explicitly incorporating asymmetric misclassification costs into the evaluation logic allows model performance to be interpreted not only in statistical terms but also in terms of tangible economic consequences, a dimension directly relevant to revenue administration objectives and the institutional accountability of customs systems.

The role of the legislative and regulatory environment also deserves explicit recognition in the context of practical implementation. Customs administrations operate within multi-layered legal frameworks that govern the permissible scope of automated decision support in enforcement activities. In many jurisdictions, administrative law requirements mandate that individual inspection decisions remain subject to human review and documented justification, irrespective of algorithmic outputs. The framework proposed in this study is consistent with these requirements: it generates a ranked prioritization list for officer review rather than automated enforcement determinations, and the SHAP-based explanations it produces provide an auditable, feature-level rationale for each risk score. Final inspection authority remains with customs officers, who continue to apply domain knowledge, consult supplementary intelligence sources, and exercise regulatory authority within the bounds of applicable law. This design principle, which positions machine learning as a decision-support tool rather than a substitute for human decision-making, is not merely a normative choice but a practical prerequisite for institutional adoption and legal compliance.

Risk scores supported by explainable artificial intelligence further strengthen institutional acceptance by making the model's internal logic transparent and contestable. Interpretability supports the integration of model-driven decisions into existing governance frameworks, including requirements for audit trails, internal oversight, and regulatory accountability. In this respect, the proposed framework is not intended as a broad policy reform but as a practical, operationally embedded methodological guideline for enhancing targeted inspection and ranking within currently deployed customs risk management systems. Taken together, these limitations do not undermine the study's core contribution; rather, this work should be understood not as proposing a definitive solution but as providing empirical evidence that time-aware evaluation, cost-sensitive decision logic, and explainable artificial intelligence can be jointly implemented in customs risk management using real import transaction data. The results provide a foundation for future research and actionable insights for improving operational risk assessment practices in customs administrations.

VI. CONCLUSION

This study empirically investigates how machine learning methods can be applied in a more operationally viable manner within real-world customs risk management settings. Using import declarations of construction materials as an empirical case, the study develops and evaluates a risk-scoring framework that integrates time-based evaluation, cost-sensitive decision logic accounting for asymmetric misclassification costs, and explainable artificial intelligence techniques. The primary objective is to assess model performance under conditions that closely reflect actual inspection environments. Throughout, the central operational problem is treated as it exists in practice: customs administrations must accurately identify high-risk declarations while operating under strictly limited inspection capacity, and any analytical framework that ignores this constraint has limited practical value, regardless of its statistical performance.

The empirical findings demonstrate that, risk score-based ranking enables more effective decision-making under constrained inspection resources. The practical advantage is direct: rather than applying a fixed decision threshold, risk-score ranking allows an administration to inspect the top K declarations by predicted risk, adjusting the inspection budget dynamically as capacity allows. Time-based evaluation enables model performance to be assessed under realistic operational conditions and provides a more reliable measure of generalization to future import declarations. In addition, incorporating cost-sensitive logic empirically demonstrates how inspection efforts can be better aligned with their underlying economic consequences, thereby supporting more effective management of revenue risk.

Results from the explainable artificial intelligence analysis further reveal that risk scores are driven by coherent interactions among multiple features, including price, weight, unit discrepancies, and behavioral patterns in import declarations. This transparency enhances the institutional interpretability of model decisions and supports the use of machine learning-based risk assessment in a manner consistent with established customs control logic rather than as opaque black-box behaviour.

Overall, the contribution of this work lies not in advancing a new theoretical paradigm but in demonstrating that time-aware evaluation, cost-sensitive ranking, and explainable AI can be jointly operationalized in a real customs environment using large-scale import transaction data. The findings provide empirical evidence that moving beyond a sole focus on predictive accuracy, toward jointly addressing inspection capacity constraints, asymmetric economic costs, and interpretability requirements, is both feasible and materially consequential for customs risk management practice.

The deployment of this framework in operational customs settings will, however, be conditioned by factors beyond model performance. Trade patterns shift in response to macroeconomic conditions, regulatory changes, and supply chain dynamics, requiring structured model review cycles as part of ongoing governance. More broadly, the legal and regulatory environment of a given customs administration shapes what automated decision-support tools can do. In most jurisdictions, individual inspection decisions must remain subject to human review and documented justification. The framework is designed with this constraint in mind: it produces ranked prioritization inputs for officer review rather than automated enforcement outcomes, and the SHAP explanations it generates serve as an auditable record of the model's reasoning. Recognizing these operational constraints is not a qualification of the study's contribution but a necessary condition for the responsible and sustainable adoption of machine learning in customs risk management.

APPENDIX A. SUPPLEMENTARY MODEL DIAGNOSTICS, FEATURE ENGINEERING, ROBUSTNESS CHECKS, AND COMPUTATIONAL EFFICIENCY

TABLE AI. PERFORMANCE DIFFERENCES BETWEEN TRAINING AND TEST SETS

Dataset Split	N (observations)	Positive Rate	AUC	AP	F1@0.5	Optimal Threshold
Training (historical)	423,374	0.3427	0.9997	0.9995	0.9884	0.447
Test (future holdout)	204,497	0.3381	0.9524	0.9191	0.8147	0.730
Difference (Test-Train)		−0.0046	−0.0473	−0.0804	−0.1737	+0.283

Note: Differences highlight the expected performance degradation under a strict future holdout evaluation, reflecting a realistic deployment scenario without information leakage.

TABLE AII. LIGHTGBM MAIN MODEL: FULL HYPERPARAMETER CONFIGURATION

Parameter	Value	Description
n_estimators	800	Number of boosting rounds
learning_rate	0.05	Step size shrinkage
num_leaves	63	Max leaves per tree
subsample	0.8	Row subsampling ratio
colsample_bytree	0.8	Column subsampling ratio
min_child_samples	20	Min data in leaf (regularization)
reg_alpha	0.0	L1 regularization
reg_lambda	0.0	L2 regularization
boosting_type	gbdt	Gradient boosting decision tree
objective	binary	Binary classification
max_depth	−1	Unlimited (controlled by *num_leaves*)

TABLE AIII. FEATURE ENGINEERING SUMMARY. ENCODING: ONE-HOT ENCODING (SKLEARN), FITTED ON TRAINING SET ONLY (HANDLE_UNKNOWN = "IGNORE")

Feature Group	Count	Details
Numeric (raw)	9	*log_qty*, *log_value_usd*, *unit_value_usd*, *tax_ratio*, *USD_exchange_rate*, *month_sin*, *month_cos*, *uv_rel_dev*, *tr_rel_dev*
Categorical (raw)	2	hs4 (HS4 product code), *Sender_country*
After OHE expansion	117	hs4 + origin country one-hot dummies
Total model inputs	126	9 numeric features plus 117 one-hot-encoded categorical dummy variables

TABLE AIV. LEARNING CURVE DATA: TRAINING AND TEST AUC BY TRAINING SET SIZE

Train Fraction	Train Size	Train AUC	Test AUC
0.10	42,337	1.0000	0.9548
0.20	84,674	1.0000	0.9536
0.35	148,180	0.9999	0.9544
0.50	211,687	0.9998	0.9506
0.65	275,193	0.9998	0.9522
0.80	338,699	0.9997	0.9531
1.00	423,374	0.9997	0.9502

Note: Training AUC ≈ 1.000 across all training sizes while Test AUC remains stable at approximately 0.950–0.955, indicating that the performance gap reflects the inherent difficulty of the temporal generalization task rather than overfitting to training data volume. The near-perfect training fit is a direct consequence of the rule-based label construction strategy: since labels are derived deterministically from the training data distribution, the model can achieve near-zero training loss without memorizing individual records. Critically, the stability of test AUC across all training sizes confirms that temporal generalization to the 2025 holdout set is consistently strong regardless of training volume, providing evidence against overfitting-driven performance inflation.

TABLE AV. PERMUTATION FEATURE IMPORTANCE ON 2025 TEST SET (N_REPEATS = 10; SCORING = AUC; NUMERIC FEATURES ONLY)

Feature	Importance (Mean AUC Drop)	Std
tr_rel_dev	0.25195	0.00079
uv_rel_dev	0.10033	0.00047
tax_ratio	0.01034	0.00009
unit_value_usd	0.00578	0.00009
month_sin	0.00109	0.00004
log_qty	0.00108	0.00003
USD_exchange_rate	0.00093	0.00004
log_value_usd	0.00058	0.00002
month_cos	-0.00047	0.00003

Note: Values represent the mean decrease in test AUC when a feature is randomly permuted. *tr_rel_dev* and *uv_rel_dev* dominates importance (0.252 and 0.100 AUC drop, respectively), confirming that the model's discriminative capacity is driven primarily by price and tax anomaly signals. *month_cos* shows a negligible negative value (−0.00047), indicating no meaningful contribution.

TABLE AVI. COMPUTATIONAL EFFICIENCY AND RUNTIME CHARACTERISTICS

Model	LightGBM
Training Time (s)	-
Inference Time (s)	12.03
Latency per 10k Records (s)	0.588
Throughput (records/s)	17,005
Model Size	0.05 MB

Note: Training time is omitted due to offline model fitting. Runtime metrics are reported for batch inference on the 2025 test set.

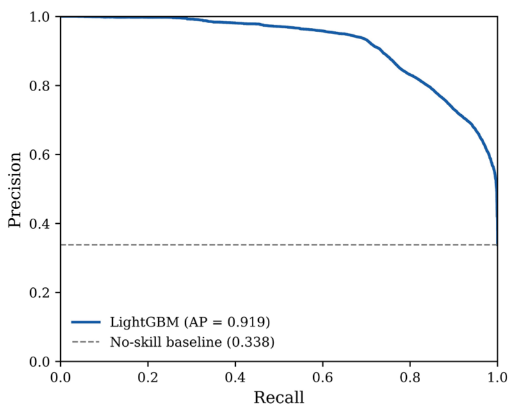

Fig. A1. Precision-recall curve (2025 test set).

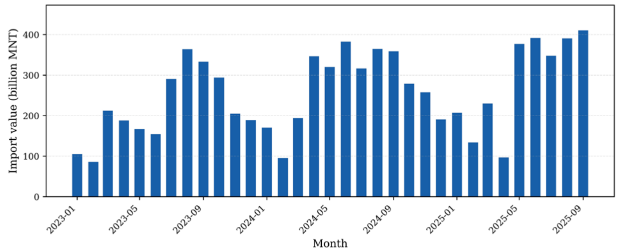

Fig. A2. Monthly dynamics of construction materials imports (Jan 2023–Sep 2025).

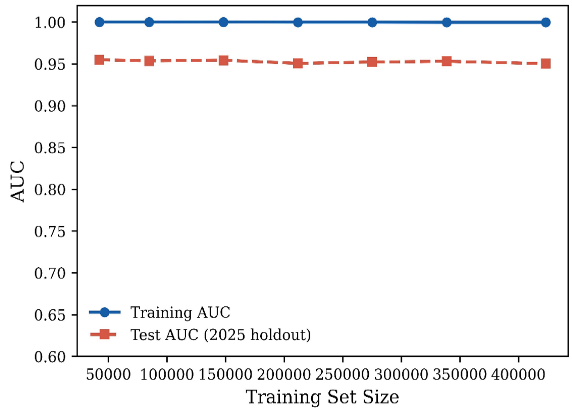

Fig. A3. Learning curves: training and test AUC as a function of training set size.

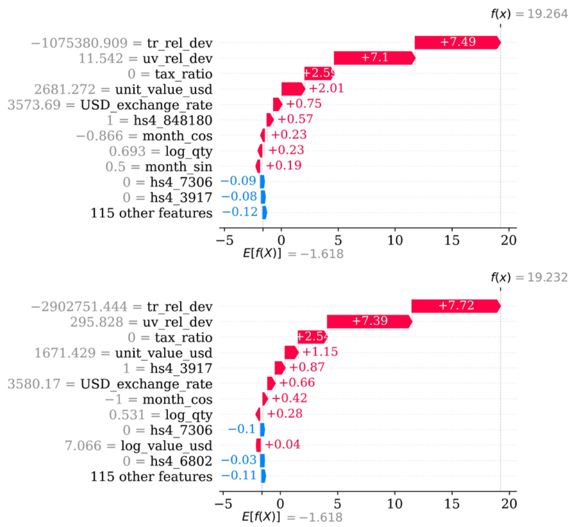

Fig. A4. Local SHAP waterfall explanation.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

T.S. conceptualized the study, developed the methodology, prepared and processed the dataset, conducted the empirical analysis, interpreted the results, and wrote the original draft. R.Y. contributed to the study design, supervised the research process, reviewed the methodology and results, and revised the manuscript critically for intellectual content. All authors reviewed and approved the final version of the manuscript.

References

[1] J. Vanhoeyveld, D. Martens, and B. Peeters, "Customs fraud detection: Assessing the value of behavioural and high-cardinality data under the imbalanced learning issue," *Pattern Analysis and Applications*, vol. 23, no. 3, pp. 1457–1477, 2020. doi: 10.1007/s10044-019-00852-w

[2] D. C. Novith, A. R. Sembiring, and M. H. Ridho, "Customs fraud detection using an extremely boosted neural network (xbnet)," *Jurnal BPPK: Badan Pendidikan Dan Pelatihan Keuangan*, vol. 16, no, 1, pp. 67–77, 2023. doi: 10.48108/jurnalbppk.v16i1.816

[3] X. Zhou, "Data mining in customs risk detection with cost-sensitive classification," *World Customs Journal*, vol. 13, no. 2, pp. 115–130, 2019. doi: 10.55596/001c.116219

[4] A. C. Bahnsen, D. Aouada, and B. Ottersten, "Example-dependent cost-sensitive decision trees," *Expert Systems with Applications*, vol. 42, no. 19, pp. 6609–6619, 2015. doi: 10.1016/j.eswa.2015.04.042

[5] R. Alwanin, M. M. B. Ismail, and O. Bchir, "Customs fraud detection using a gradient boosting approach for joint classification and risk estimation," *Scientific Reports*, vol. 16, 3432, 2025. doi: 10.1038/s41598-025-33382-z

[6] C. Nelson, "Machine learning for detection of trade in strategic goods: An approach to support future customs enforcement and outreach," *World Customs Journal*, vol. 14, no. 2, 2020. doi: 10.55596/001c.116422

[7] K. Singh, Y. C. Tsai, C. T. Li *et al.*, "GraphFC: Customs fraud detection with label scarcity," in *Proc. the 32nd ACM International Conference on Information and Knowledge Management*, 2023, pp. 4829–4835. doi: 10.1145/3583780.3614690

[8] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint*,* arXiv:1702.08608, 2017.

[9] M. Liao, W. Jiao, and J. Zhang, "Research on trade credit risk assessment for foreign trade enterprises based on explainable machine learning," *Information*, vol. 16, no. 10, 831, 2025. doi: 10.3390/info16100831

[10] World Customs Organization. (2003). Risk management guide. WCO safe package: WCO tools to secure and facilitate global trade. [Online]. Available: https://www.wcoomd.org/en/topics/facilitation/instrument-and-tools/frameworks-of-standards/safe_package.aspx

[11] World Customs Organization. (2018). WCO customs risk management compendium (vols. 1–2). [Online]. Available: https://www.wcoomd.org/en/Topics/Facilitation/Instrument%20and%20Tools/Tools/Risk%20Management%20Compendium

[12] W. Yaqin and S. Yuming, "Classification model based on association rules in customs risk management application," in *Proc. 2010 International Conference on Intelligent System Design and Engineering Application*, 2010, vol. 1, pp. 436–439. doi: 10.1109/isdea.2010.276

[13] B. Chermiti, "Establishing risk and targeting profiles using data mining: Decision trees," *World Customs Journal*, vol. 13, no. 2, pp. 39–58, 2019. doi: 10.55596/001c.116213

[14] D. J. Hand, "Measuring classifier performance: A coherent alternative to the area under the ROC curve," *Machine Learning*, vol. 77, no. 1, pp. 103–123, 2009. doi: 10.1007/s10994-009-5119-5

[15] J. Kavoya, "Machine learning for intelligence driven customs management," *African Tax and Customs Review*, vol. 1, no. 3, pp. 50–58, 2020.

[16] D. Widdowson, "Managing customs risk and compliance: An integrated approach," *World Customs Journal*, vol. 14, no. 2, pp. 63–80, 2020. doi: 10.55596/001c.116418

[17] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. doi: 10.1023/a:1010933404324

[18] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001. doi: 10.1214/aos/1013203451

[19] V. Gancheva, G. Popov, O. Nakov *et al.*, "Customs fraud detection method based on artificial intelligence and data analytics," in *Proc. 2024 4th International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, 2024, pp. 1–6. doi: 10.1109/iceccme62383.2024.10796800

[20] S. A. Sa'adah and A. Hartanto, "Fraud detection using data analytics: A case study of under-invoicing importation fraud in Indonesia," *Jurnal BPPK: Badan Pendidikan Dan Pelatihan Keuangan*, vol. 16, no. 1, pp. 146–153, 2023. doi: 10.48108/jurnalbppk.v16i1.823

[21] D. A. N. Seck, "Building machine learning models for fraud detection in customs declarations in Senegal," *WSEAS Transactions on Information Science and Applications*, vol. 21, pp. 208–215, 2024. Portico. doi: 10.37394/23209.2024.21.20

[22] R. Triepels, H. Daniels, and A. Feelders, "Data-driven fraud detection in international shipping," *Expert Systems with Applications*, vol. 99, pp. 193–202, 2018. doi: 10.1016/j.eswa.2018.01.007

[23] Y. S. Choi, "Identifying trade mis-invoicing through customs data analysis," *World Customs Journal*, vol. 13, no. 2, pp. 59–76, 2019. doi: 10.55596/001c.116214

[24] M. Spichakova and H. M. Haav, "Using machine learning for automated assessment of misclassification of goods for fraud detection," in *Proc. International Baltic Conference on Databases and Information Systems*, 2020, pp. 144–158. doi: 10.1007/978-3-030-57672-1_12

[25] B. Chan, I. Ng, and N. Chung, "Anomaly detection in trade declarations using deep learning techniques: A risk-assessment approach to identify misclassification and incorrect valuation," *Statistical Journal of the IAOS*, vol. 39, no. 4, pp. 825–834, 2023. doi: 10.3233/sji-230081

[26] W. Eberle and L. Holder, "Anomaly detection in data represented as graphs," *Intelligent Data Analysis*, vol. 11, no. 6, pp. 663–689, 2007. doi: 10.3233/ida-2007-11606

[27] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, 2009. doi: 10.1145/1541880.1541882

[28] M. M. Breunig, H. P. Kriegel, R. T. Ng *et al.*, "LOF: Identifying density-based local outliers," in *Proc. ACM SIGMOD International Conference on Management of Data*, 2000, pp. 93–104. doi: 10.1145/335191.335388

[29] F. T. Liu, K. M. Ting, and Z. H. Zhou, "Isolation forest," in *Proc. 2008 Eighth IEEE International Conference on Data Mining*, 2008, pp. 413–422. doi: 10.1109/icdm.2008.17

[30] G. Pang, C. Shen, L. Cao *et al.*, "Deep learning for anomaly detection: A review," *ACM Computing Surveys*, vol. 54, no. 2, pp. 1–38, 2021. doi: 10.1145/3439950

[31] D. T. Oosterman, W. H. Langenkamp, and E. L. van Bergen, "Customs risk assessment based on unsupervised anomaly detection using autoencoders," in *Proc. SAI Intelligent Systems Conference*, 2021, 668–681. doi: 10.1007/978-3-030-82193-7_45

[32] N. V. Chawla, K. W. Bowyer, L. O. Hall *et al.*, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002. doi: 10.1613/jair.953

[33] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429–449, 2002. doi: 10.3233/IDA-2002-6504

[34] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009. doi: 10.1109/tkde.2008.239

[35] P. T. Mazumder, "Data Driven Detection of Trade-Based Money Laundering (TBML): A predictive analytics framework for securing US supply chains and financial integrity," *SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology*, vol. 15, no. 4, pp. 423–432, 2023. doi: 10.18090/samriddhi.v15i04.07