Predictive VAT Non-Compliance Benchmarking Across Industries in Rwanda: A Machine Learning Approach

Celestin Niyomugabo[1,2,3]; Sunday A. Idowu[1]

[1]V-Analytics Department, VONSUNG, Kigali, Rwanda
[2]Department of Strategy and Risk Analysis, Rwanda Revenue Authority, Kigali, Rwanda
[3]Adventist University of Central Africa, Kigali, Rwanda

Publication Date: 2025/10/11

Abstract: Value Added Tax (VAT) non-compliance remains a persistent challenge in Rwanda despite the nationwide rollout of Electronic Billing Machine (EBM) and other digital reforms. While retrospective VAT gap studies have been useful in quantifying the scale of revenue loss, they fall short of providing predictive insights that can proactively prevent non-compliance. To address this gap, this study developed and validated an industry-aware machine learning model capable of predicting VAT non-compliance using integrated administrative microdata. The study also benchmarked VAT non-compliance across taxpayer scales and industries to identify systematic sectoral heterogeneity to generate actionable evidence for risk-based auditing and more targeted policy design. This study integrated VAT declarations, EBM transactions, and customs import records from Rwanda Revenue Authority (RRA) for the period 2020–2024, linking them at the taxpayer level to build a comprehensive compliance dataset. An Extreme gradient Boosting (XGBoost) classifier was applied, with class imbalance addressed through weighting to ensure that the minority class of VAT non-compliant returns contributed proportionately to model learning. Hyperparameters were optimized through grid search and validation to ensure robust generalization, while decision thresholds were tuned to prioritize high recall without compromising precision. Model performance was evaluated using accuracy, precision, recall, F1-score, and both ROC-AUC and PR-AUC, with additional out-of-time validation to confirm stability. Feature interpretability was ensured through SHARP-based importance analysis, which highlighted the relative contribution of discrepancies between EBM sales and declared turnover, penalty history, and trade activity in predicting VAT non-compliance. The model achieved high predictive performance for the non-compliant class (accuracy 98.9%, precision 0.932, recall 0.887, F1-score 0.909) with robust generalization across tax years. The VAT non-compliance is 6.9% overall, with statistically significant between-industry dispersion (ANOVA p-value<0.001). Elevated risk appears in transport and storage, wholesale and retail trade, manufacturing, mining and quarrying, electricity, gas, steam & air conditioning supply, and activities of households as employers. Non-compliance also increases with taxpayer scale (large 11.5%, medium 9.4%, small 6.0%). Feature importance confirms the operational salience of EBM sales and total value of supplies declared discrepancies and penalty history.

- ***Conclusion:***

Industry-aware predictive analytics can materially strengthen risk-based auditing in Rwanda by targeting higher-risk sectors and scales, improving audit efficiency and revenue recovery, and providing replicable benchmarks for sector-specific policy design.

Keywords: *VAT, Non-Compliance, Machine Learning, Rwanda, EBM.*

How to Cite: Celestin Niyomugabo; Sunday A. Idowu (2025) Predictive VAT Non-Compliance Benchmarking Across Industries in Rwanda: A Machine Learning Approach. *International Journal of Innovative Science and Research Technology*, 10(10), 507-519. https://doi.org/10.38124/ijisrt/25oct139

I. INTRODUCTION

This template, Value Added Tax (VAT) is a cornerstone of modern tax systems and one of the most significant sources of government revenue globally [1]. In many developing economies, VAT contributes between a quarter and a third of total tax collections, serving as a stable and broad-based revenue instrument that underpins fiscal sustainability [2] [3] [4]. Its attractiveness lies in its design as a multi-stage consumption tax levied on value addition at each stage of production and distribution, thereby spreading the tax burden across supply chains while limiting cascading effects [5] [6] [7]. However, the effectiveness of VAT as a revenue mobilization tool depends fundamentally on taxpayer compliance. Non-compliance, whether through under-reporting of sales, false claims of input credits, or failure to file returns, erodes revenue potential and undermines the equity and efficiency of the tax system [8].

Rwanda provides a compelling case for examining VAT compliance challenges. Since the establishment of Rwanda Revenue Authority (RRA) in 1998, the government has invested heavily in tax modernization, including electronic filing systems, online payment platforms, and the nationwide rollout of Electronic Billing machine (EBM) in 2013 [9] [10]. EBMs were introduced to automate invoice issuance and reduce opportunities for under-reporting by providing real-time, transaction-level records of taxable sales. While these reforms have significantly enhanced transparency and data availability, persistent compliance gaps remain evident. Studies and administrative reviews have shown that taxpayers continue to engage in practices such as misreporting taxable transactions, and under-declaration of sales [8] [11]. This persistence suggests that while digitalization improves monitoring capacity, it does not automatically translate into compliance unless coupled with advanced analytical and enforcement strategies [12].

Traditional approaches to VAT non-compliance measurement in Rwanda and across Sub-Saharan Africa have been largely retrospective. Top-down VAT gap analyses which compare theoretical VAT liabilities against actual collections, provide estimates of revenue loss but cannot identify which taxpayers or sectors are responsible. Similarly, audit-based studies offer insights into evasion practices but remain resource-intensive and limited in scope, covering only a fraction of the taxpayer population. These methods are insufficient for proactive non-compliance management in a context where administrative capacity is constrained and enforcement resources are scarce. As a result, many high-risk taxpayers remain undetected, and compliance efforts often rely on random or broad-based audits that may not yield optimal results.

Emerging international practice points to the potential of data-driven, risk-based approaches. Tax administrations in high-income countries have increasingly applied advanced analytics and machine learning to taxpayer data to predict fraud risk, detect anomalies, and prioritize audits. For instance, a study in Italy found that replacing the 10% least-effective audits with ml-selected cases could have increased recovered tax evasion by about 38% [13]. South African Revenue Service (SARS) has also been a pioneer in Machine Learning (ML) risk profiling, with about one-third of its compliance revenue being attributed to automatic risk-scoring using ML on diverse data sources [14]. However, the application of machine learning to tax compliance generally remains limited in Sub-Saharan Africa, Rwanda included.

Despite global advances in economic development and Rwanda's significant investment in electronic tax systems, concerns over VAT non-compliance remain and demand research that moves beyond descriptive assessments toward predictive modeling in the Rwandan context. To date, most analyses of VAT non-compliance in Rwanda have been descriptive, focusing on measuring compliance gaps after they occur rather than anticipating them through predictive approaches. To address this gap, this study focused on developing an industry-aware machine learning predictive model for VAT non-compliance across industries in Rwanda. It relied on administrative datasets obtained from RRA, which were pre-processed and harmonized for analysis. Descriptive and exploratory analyses were conducted to examine VAT non-compliance behavior and its distribution across industries, followed by the development and evaluation of an ML model tailored to industry-specific compliance risks. In addition, a user-friendly interface was designed and implemented to enable stakeholders to interact with the predictive model and explore non-compliance risk insights. Finally, the study developed actionable, data-driven policy recommendations to strengthen VAT compliance across industries in Rwanda, drawing on the model outputs and sectoral benchmarks.

II. METHODOLOGY

➢ *Data Collection*

This study relied exclusively on secondary data obtained from RRA's administrative databases. The audit data which represent the outcome of past VAT audit assessments and VAT returns data were extracted from RRA's E-Tax system. The EBM transaction data which consist of invoice-level sales information captured in real time were accessed through EBM system. This data source provided detailed records of taxable supply activities across firms. Additionally, customs import data were sourced from RRA's Automated System for Customs Data (ASYCUDA) system.

➢ *Research Design*

The This study adopted a quantitative, explanatory research design aimed at forecasting patterns of VAT non-compliance across industries in Rwanda using ML techniques. The purpose of this design is to identify statistical relationships between firm-level attributes and the likelihood of VAT non-compliance. To achieve this, the study employed a supervised machine learning framework formulated as a binary classification task, where historical administrative tax data serve as labeled training data. The model classifies taxpayers as compliant (0) or non-compliant (1) based on their feature profiles and generates predictive outputs that can be applied to future cases for proactive compliance management.

Among the available machine learning algorithms, this study employed Extreme gradient Boosting (XGBoost) because it has several advantages for the VAT non-compliance task. XGBoost is a tree-ensemble method that consistently achieves high predictive accuracy on structured datasets and has outperformed traditional approaches such as logistic regression, single decision trees, and even random forests in similar fraud detection and tax compliance studies [15]. It also heterogeneous administrative data effectively, accommodating both numerical and categorical variables without the need for complex preprocessing. In addition, XGBoost integrates L1 (Lasso) and L2 (Ridge) regularization, which mitigates overfitting in noisy datasets where patterns may be influenced by a few dominant variables. The XGBoost algorithm is highly efficient with large datasets, with scalability and parallel processing capabilities that are essential when working with millions of EBM transaction records spanning multiple fiscal years. Moreover, because of its tree-based architecture, XGBoost demonstrates robustness to outliers, an important property in tax data where some taxpayers report exceptionally high turnover values that would otherwise distort predictions.

➢ *Taxpayer Selection Criteria*

This study targeted VAT-registered taxpayers in Rwanda who have transacted using EBM during the period 2020-2024, including both monthly and quarterly VAT filers. The population covered businesses of various scales (small, medium, and large) and across diverse ISIC sections, including those involved in domestic sales and international trade.

To ensure consistency, completeness, and data integrity, the following categories were excluded from the study:

- Entities lacking EBM transaction data, including informal businesses not registered with RRA, VAT-registered taxpayers without EBM devices, and businesses not registered for VAT.
- Industries structurally outside VAT benchmark such as agriculture, forestry, and fishing due to VAT exemptions on agricultural products.
- Industries which are not required to issue EBM receipts due to the nature of their operations. These industries include financial and insurance services, and education.
- Public or non-commercial institutions that are not liable for VAT including government agencies, compulsory social security entities, and human health or social work activities.

➢ *Data Preprocessing, eda, and Feature Engineering*

Prior to model development, the data underwent extensive preprocessing and exploratory analysis to ensure quality and suitability. Missing values, duplicates, inconsistent formats, and outliers were addressed using Python's Pandas library, while XGBoost's sparsity-aware property was leveraged to handle remaining gaps without explicit imputation. Exploratory Data Analysis (EDA) examined feature distributions, correlations, and sectoral compliance patterns using visualizations such as histograms, boxplots, and heatmaps, and statistical tests including chi-square for categorical associations.

Feature selection combined domain knowledge and data-driven methods, retaining variables shown to be significant in prior VAT compliance research, consistently available across the dataset, and relevant to taxpayer behavior. Multicollinearity was controlled by dropping or consolidating highly correlated variables, while preliminary information gain rankings from early XGBoost runs guided retention of features with predictive value. Feature engineering enhanced the dataset through standardization of column names, filtering of the VAT registry to include only active taxpayers with defined filing obligations, and creation of new variables including missed declarations and a late-filing flag. EBM transaction data were aggregated to capture invoice counts, sales, and VAT per period and merged to returns, while customs records were integrated by TIN to derive import-to-sales ratios. Finally, continuous variables were normalized, categorical variables label-encoded, and the dataset harmonized to provide a robust input structure for the predictive model.

➢ *Target Variable and Modeling Approach*

The target variable was constructed as a binary indicator representing VAT non-compliance status, where a value of 1 denotes non-compliance and 0 denotes compliance. This classification based primarily on historical audit outcomes obtained from RRA's administrative records. Firms flagged through prior audits as having underreported VAT liabilities, misstated sales, or engaged in invoice suppression were labeled as non-compliant. XGBoost was used to predict the binary target due to its high accuracy, ability to handle missing values, and robustness to multicollinearity. XGBoost's gradient boosting framework builds an ensemble of decision trees sequentially, where each new tree attempts to correct the errors of its predecessors. This makes it particularly well-suited for VAT non-compliance modeling, where feature interactions are often non-linear and the dataset may contain complex behavior patterns.

From the training data $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^{n}$, the target of a given VAT declaration i is expressed as follows:

$$y_i = \begin{cases} 1, & \text{if taxpayer } i \text{ is non-compliant,} \\ 0, & \text{if taxpayer } i \text{ is compliant.} \end{cases} \qquad y_i \in \{0,1\}$$

➢ *Model Specification*

Implementing XGBoost classifier optimized a regularized binary logistic loss. The model outcome is expressed as:

$y_i \in \{0,1\}$: Binary indicator of compliance for taxpayer i

Where:

- $y_i = 1$: Non-compliant
- $y_i = 0$: Compliant

$\hat{p}_i = P(y_i = 1 \mid \mathbf{X}_i)$: Probability of non-compliance given feature vector $\mathbf{X}_i$

The model minimizes the regularized binary logistic loss:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} [-y_i \log(\hat{p}_i) - (1 - y_i)\log(1 - \hat{p}_i)] + \Omega(\phi)$$

Where:

- ϕ denotes the set of all decision trees,
- $\Omega(\phi)$ is a regularization term penalizing tree complexity to prevent overfitting.

➢ *Model Training and Evaluation*

The predictive model was trained by partitioning the dataset into training and test sets using stratified sampling to preserve the distribution of compliant and non-compliant taxpayers, with 20% reserved as a held-out test set. From the remaining 80% training portion, a further 15% was set aside as validation data to support early stopping and mitigate overfitting, leaving roughly 68% of the data for model fitting. The model was implemented using the XGBoost classifier with a logistic objective optimized for binary classification, and class imbalance was addressed through the scale_pos_weight parameter, calculated as the ratio of negatives to positives in the training set to improve sensitivity to the minority non-compliant class.

Hyperparameters were tuned to balance predictive accuracy and generalization, with the final specification including 600 boosting rounds, a maximum tree depth of 4, a learning rate of 0.05, subsampling ratios of 0.8 for rows and 0.6 for columns, and regularization terms set $\lambda = 6.0$ (L2) and $\alpha = 2.0$ (L1), together with a minimum child weight of 5 and a gamma value of 1.0 to control tree complexity; the hist tree method was used to improve computational efficiency with large-scale EBM data.

Model performance was monitored on the validation set using area under the precision–recall curve (AUC–PR) as the principal evaluation metric, with early stopping applied after 50 rounds without improvement, and the best iteration retained for final testing. Evaluation on the held-out test set was based on probabilistic outputs transformed into class labels using the default 0.5 threshold unless alternative cut-offs were defined by operational constraints; robustness was examined using tuned thresholds that maximized F1 or maximized precision subject to a recall floor. Performance was summarized through the confusion matrix and derived metrics of precision, recall, and F1-score, with AUC–PR and AUC–ROC used as ranking metrics; emphasis was placed on AUC–PR given the imbalanced data.

To confirm generalization and guard against overfitting, additional checks included three-fold stratified cross-validation on the training set to assess variance across resamples, learning curves for AUC–PR and ROC–AUC to monitor convergence patterns as the training sample increased, and out-of-time validation in which the model was trained on earlier tax periods and tested on later ones to directly evaluate temporal stability.

III. IMPLEMENTATION, RESULTS, INTERPRETATION, AND DISCUSSION OF FINDINGS

A. Profile of VAT Payers

Across the five-year period, the integrated administrative record shows a large and active VAT base. A total of 987,467 VAT declarations were identified for 55,936 VAT-registered taxpayers, of which 37,174 were outside the analytical scope defined in this study. The remaining 950,293 valid returns constitute the empirical basis for measuring compliance outcomes. In addition to VAT return records, the study also examined a total of 275,324,272 EBM transaction data for aggregated into 1,984,562 EBM transaction summaries and 243,517 import declaration records covering the same period, which were used in later stages for cross-referencing and discrepancy analysis.

Generally, the number of VAT payers increased over time from 2020 to 2023 with a decline in 2024 which is linked with RRA’s effort to de-register non-active taxpayers [4]. The trend in total turnover increased overtime from 6,513.9 bn in 2020 to 21,687.9 bn in 2024.

Although the volume of declarations necessarily fluctuates with firm entry/exit and filing frequency rules, the size of the analytic base provides high statistical power to detect differences across scale and industry, and to train a predictive model. The filing frequency is strongly stratified by taxpayer scale, which influences both observability and compliance incentives. Among small taxpayers, 71.7% are registered to file quarterly, implying a maximum of four returns per calendar year; by contrast, the quarterly regime is used by only 2.2% of medium taxpayers and 0.9% of large taxpayers. The expected filing burden for medium and large segments is thus predominantly monthly, aligning with their larger transaction volumes and administrative capacity.

Taken over the 2020–2024 horizon, this stratification yields markedly different observed filing counts per taxpayer. Large taxpayers submitted 12,436 returns across 213 entities (mean 58.4 returns per taxpayer over the period), medium taxpayers submitted 28,682 returns across 495 entities (mean 57.9), while small taxpayers submitted 909,175 returns across 53,258 entities (mean 17.1). The near-sixty-return averages for large and medium segments indicate consistently high filing regularity under a monthly regime, whereas the smaller average for the small segment is mechanically consistent with quarterly filing and higher churn in the small-taxpayer population.

These patterns show that Rwanda’s VAT system exhibits high baseline compliance with filing obligations among medium and large taxpayers as proxied by the high number of observed filings per entity, yet this does not automatically translate into superior substantive compliance of accurate self-assessment, as shown in Table 1. They also show that small taxpayers’ lighter filing burden and simpler operations may constrain the scope for complex misreporting but can interact with monitoring gaps.

Table 1 VAT Filing Behavior by Business Scale

Scale	**Taxpayers**	**VAT returns (2020-2024)**	**Number of VAT returns**[1]
Large	213	12,436	58.4
Medium	495	28,682	57.9
Small	53,258	909,175	17.1

[1] Average number of VAT returns submitted by a taxpayer from 2020 to 2024

B. Descriptive Analysis and VAT Non-Compliance Across Industries

The correlation analysis (see Figure 1) highlights two main clusters: customs-related variables such as import value, customs VAT record, customs activity volume, and import VAT contribution show strong positive associations, indicating internal consistency within the import channel; while on the sales side, recorded sales correlate notably with zero-rated sales and, to a lesser extent, with import values and invoice value index. By contrast, declared turnover shows only weak alignment with EBM-based sales and activity counts, suggesting a disconnect between taxpayer declarations and electronically recorded sales. Taken together, these patterns confirm that the dataset contains internally coherent variable blocks while also preserving non-redundant variation across clusters and highlight a balance that strengthens the predictive power of the model and contributes positively to its high accuracy.

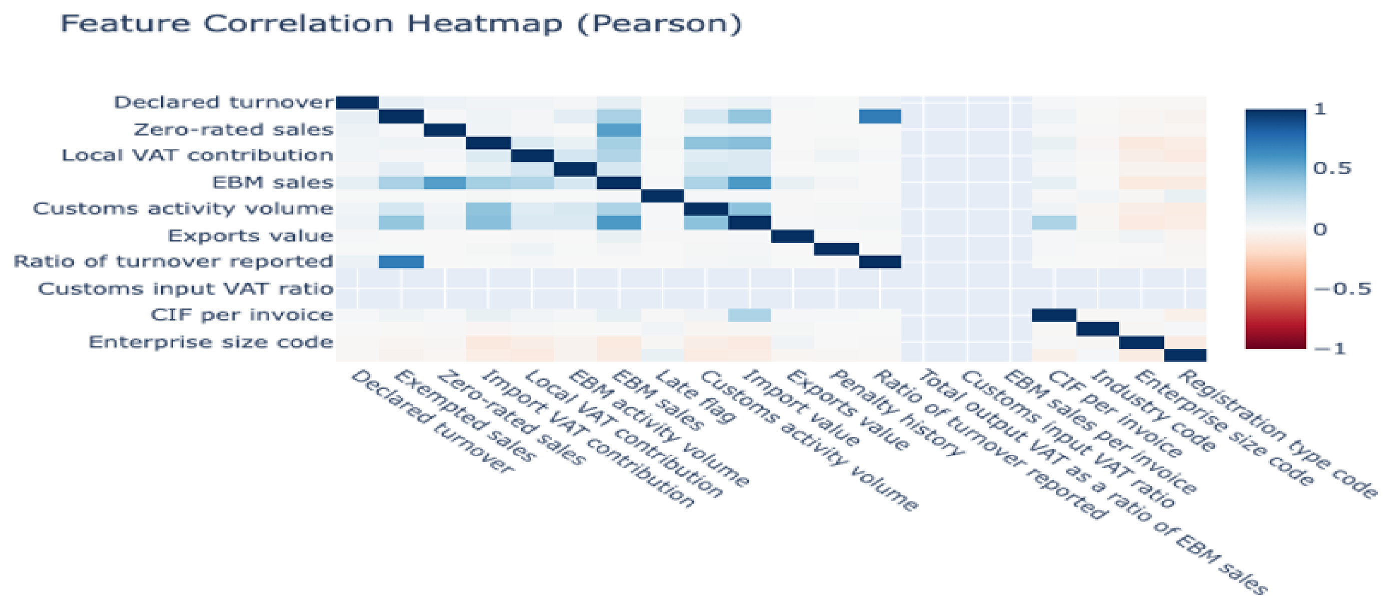

Fig 1 Pearson Correlation Between Variables

Descriptive statistics (Table 2) show large statistically significant level differences between vat compliant and non-compliant returns across every variable (all $p < 0.001$). Non-compliant filings are consistently larger: mean total value of supplies 133.8 Bn vs. 59.8 Bn for compliant; exempted sales 46.7 vs. 15.1; zero-rated sales 7.7 vs. 5.9; exports 2.3 vs. 0.9; reported output vat 13.9 vs. 6.8; customs input vat 2.4 vs. 0.9; local input vat 5.8 vs. 3.1; and ebm sales 166 vs. 50.6 (all in bn). Medians underscore heavy right-skew, yet still higher for the non-compliant (e.g., total supplies median 6.3 Bn vs. 1.5 Bn; EBM sales median 18.8 Bn vs. 0).

A notable pattern is the declaration–EBM contrast. Among compliant returns, declared sales slightly exceed recorded EBM sales (59.8 vs. 50.6 Bn), consistent with timing and legitimate adjustments; among non-compliant returns, declared sales fall materially below EBM sales (133.8 vs. 166 Bn), aligning with under-reporting behavior. This clear separation in levels and medians provides strong signal for the classifier and helps explain the model's high accuracy as high-value and high-discrepancy cases are systematically associated with non-compliance, making them easier to detect.

Table 2 Descriptive Statistics

VAT non-compliant	Count	Mean - bn	std[2] - bn	Median - bn	P-value
Total Value of Supplies					
No	891,402	59.8	5,572.5	1.5	0.000
Yes	58,888	133.8	1,061.5	6.3	
Total	950,290	64.4	5,403.5	1.7	
Exempted sales					
No	891,402	15.1	442.3	0	0.000
Yes	58,888	46.7	731.7	0	
Total	950,290	17.1	465.6	0	
Zero-rated sales					
No	891,402	5.9	390.3	0	0.000
Yes	58,888	7.7	152.8	0	
Total	950,290	6	380	0	
Export					

[2] Standard Deviation

VAT non-compliant	Count	Mean - bn	std^2 - bn	Median - bn	P-value
No	891,402	0.9	42.7	0	0.000
Yes	58,888	2.3	89.7	0	
Total	950,290	1	47	0	
Reported output VAT					
No	891,402	6.8	997	0.1	0.000
Yes	58,888	13.9	118.8	0.5	
Total	950,290	7.2	966.1	0.1	
Reported customs input VAT					
No	891,402	0.9	11.6	0	0.000
Yes	58,888	2.4	30	0	
Total	950,290	1	13.5	0	
Reported local input VAT					
No	891402	3.1	28.4	0	0.000
Yes	58888	5.8	42.5	0.3	
Total	950290	3.2	29.5	0	
EBM sales					
No	891,405	50.6	545	0	0.000
Yes	58,888	166	1,221.5	18.8	
Total	950,293	57.8	609.8	0.4	

C. VAT Non-Compliance by Taxpayer Scale

A key result of the analysis is that substantive VAT non-compliance is not lowest at the top of the size distribution (Table 3). Using the binary non-compliance outcome defined in the analytic dataset, large taxpayers exhibit an 11.5% non-compliance rate, medium taxpayers 9.4%, and small taxpayers 6.0%. The gradient is unambiguous: larger entities have higher measured non-compliance than smaller entities despite their higher administrative capacity and higher filing regularity.

Table 3 VAT Non-Compliance by Business Scale

Scale	VAT non-compliance
Large	11.5
Medium	9.4
Small	6.0

D. Benchmarking VAT Non-Compliance Across Industries

- ➢ *National Baseline and Sectoral Dispersion*

The overall VAT non-compliance rate over 2020–2024 is estimated at 6.9%, calculated in reference to the VAT returns that are biased out of all submitted VAT declarations. This national baseline masks substantial heterogeneity across industries, as classified by ISIC sections. Above-average VAT non-compliance is observed in activities of households as employers; transport and storage; other service activities; mining and quarrying; wholesale and retail trade; electricity, gas, steam and air-conditioning supply; and manufacturing. The dispersion is economically meaningful as the sectors characterized by complex supply chains, cash-intensive transactions, or extensive zero-rating tend to exhibit higher VAT non-compliance.

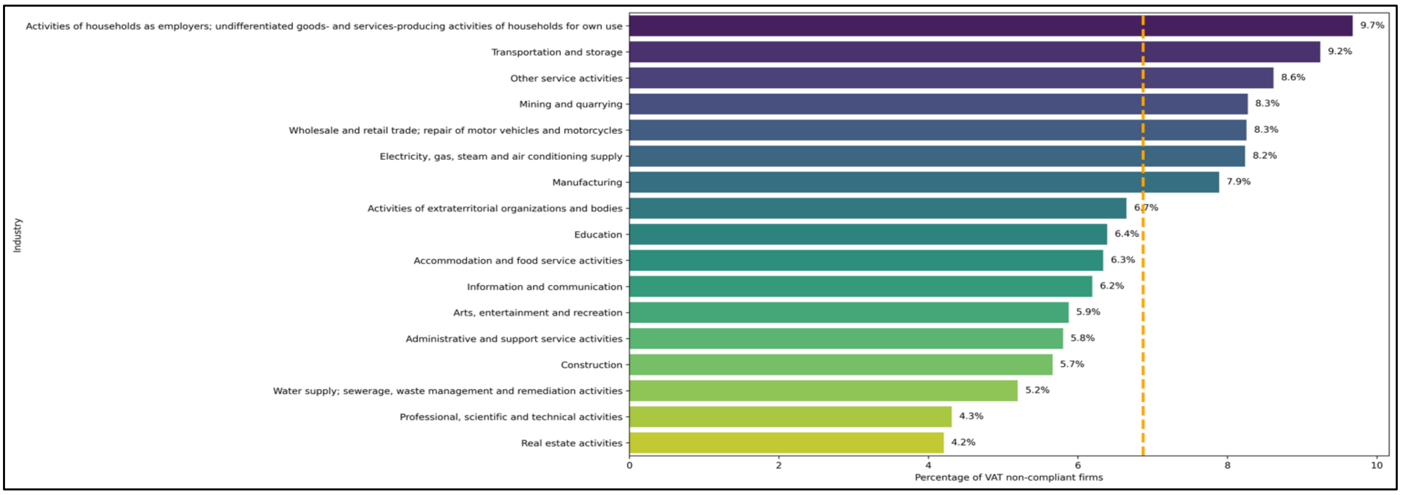

Fig 2 VAT Non-Compliance by Industry

- *Statistical Significance of Between-Industry Differences*

The statistical testing confirms that VAT non-compliance rates differ systematically across industries. A one-way ANOVA rejects the null hypothesis of equal mean VAT non-compliance rates across International Standard of Industrial Classification (ISIC) sections, with a p-value reported as p<0.001. Post-hoc multiple comparisons (Tukey HSD) identify numerous pairs of industries exhibiting statistically significant differences, with several comparisons yielding extremely small p-values. The weight of evidence indicates that sectoral patterns are not due to sampling noise in this large administrative dataset; rather, there are stable structural differences across industries that are relevant for policy design.

For interpretative clarity, two points are emphasized. First, statistical significance at very small p-values in large samples is unsurprising, but here it is accompanied by substantively large gaps relative to the 6.9% national baseline in several sectors, justifying practical policy differentiation. Second, the multiplicity-adjusted post-hoc results demonstrate that the pattern is not driven by a single outlier sector; instead, clusters of higher-risk sectors stand out against lower-risk comparators. The post-hoc test results can be found from Table 4.

Table 4 Post-hoc Analysis of VAT Non-Compliance Difference Across Industries

Industry 1	Industry 2	p-value		Industry 1	Industry 2	p-value		Industry 1	Industry 2	p-value
J	G	4.00E-39		C	R	1.00E-10		N	D	1.36E-10
J	C	6.00E-22		F	M	5.00E-30		N	L	7.41E-11
J	M	6.00E-35		F	I	9.00E-07		N	B	5.14E-13
J	S	4.00E-31		F	S	2.00E-74		N	T	0.00018453
J	H	7.00E-45		F	P	3.00E-05		S	P	2.90E-21
J	D	6.00E-08		F	H	4.00E-100		S	E	9.28E-14
J	L	2.00E-16		F	D	5.00E-14		S	L	8.71E-58
J	B	6.00E-10		F	L	1.00E-11		S	R	9.38E-16
G	F	9.00E-156		F	B	5.00E-18		P	H	1.48E-31
G	M	6.00E-224		F	T	7.00E-05		P	D	2.98E-06
G	I	2.00E-41		M	I	1.00E-44		P	L	1.05E-17
G	N	1.00E-40		M	N	3.00E-19		P	B	1.25E-07
G	P	5.00E-23		M	S	4.09E-143		H	E	1.49E-17
G	H	2.00E-08		M	P	7.74E-34		H	L	7.09E-70
G	E	2.00E-12		M	H	3.30E-173		H	R	1.45E-21
G	L	1.00E-60		M	D	6.88E-36		D	E	1.45E-08
G	R	1.00E-14		M	B	1.59E-45		D	L	1.35E-26
C	F	9.00E-71		M	R	2.55E-10		D	R	2.16E-07
C	M	5.00E-143		M	T	1.97E-09		E	B	2.15E-09
C	I	2.00E-21		I	S	9.82E-31		E	T	3.03E-05
C	N	5.00E-26		I	H	2.45E-45		L	B	4.20E-31
C	S	1.00E-04		I	D	3.82E-07		L	R	4.59E-08
C	P	2.00E-13		I	L	2.63E-19		L	T	2.17E-09
C	H	1.00E-11		I	B	6.00E-09		B	R	1.67E-08
C	E	4.00E-10		N	S	4.12E-35				
C	L	2.00E-48		N	H	4.85E-48				

- *Sector Profiles for Higher-Risk Industries*

Sectoral analysis highlights distinct compliance risks across higher-risk industries. In transport and storage, above-baseline VAT non-compliance reflects operational complexities such as cash-based last-mile logistics, ancillary fees, and subcontracting layers that weaken audit trails, with EBM-linked ratios proving most predictive of discrepancies. In wholesale and retail trade, risks persist despite EBM coverage, particularly in allocating sales between taxable, zero-rated, and exempt categories, and in input VAT claims on fast-turnover inventories. Manufacturing shows elevated risk due to multi-stage production, extensive input VAT credits, and classification choices that significantly affect liability, a finding corroborated by the model's emphasis on ratios of non-taxable to total EBM sales. For utilities (electricity, gas, steam, and air-conditioning), divergences stem largely from tariff classification and rate application rather than concealed turnover. In mining and quarrying, risks emerge from export zero-rating, large input VAT refunds, and timing strategies linked to long refund cycles and project-based accounting.

Finally, other service activities and households as employers display high variance in behavior due to weaker EBM coverage and episodic revenue streams, where administrative challenges outweigh deliberate evasion.

E. Predictive Modelling Results

➢ *Overall Performance*

XGBoost was trained on the integrated administrative dataset to identify returns at heightened risk of non-compliance. As shown in Table , on a held-out test set comprising 190,059 observations (178,281 compliant and 11,778 non-compliant), the classifier achieved an overall accuracy of 0.989, with balanced class performance despite marked imbalance. For the policy-relevant non-compliant class, precision is 0.932 and recall is 0.887, yielding an F1-score of 0.909; thus, approximately 93% of flagged returns are truly non-compliant and the model identifies about 89% of all non-compliant cases. The false-alarm rate on compliant returns is correspondingly low (0.43%), indicating efficient use of audit capacity. While the weighted averages converge to 0.989 which reflects the dominance of the compliant class, the macro-average F1 of 0.951 confirms strong performance across both classes rather than accuracy being driven solely by base rates.

Table 5 Overall Performance of the Model

Class	Precision	Recall	F1-score	Support
Compliant	0.993	0.996	0.994	178,281
Non-compliant	0.932	0.887	0.909	11,778
accuracy			0.989	190,059
macro avg	0.962	0.941	0.951	190,059
weighted avg	0.989	0.989	0.989	190,059

Model performance accuracy remained consistent across different evaluations. To further validate generalization, an out-of-time (OOT) validation was conducted by training the model on earlier periods and testing it on later tax period years. The results demonstrated comparable accuracy and stability, confirming that the model maintains predictive strength across time, as summarized in Table 6.

Table 6 Overall OOT Performance of the Model

Class	Precision	Recall	F1-score	Support
Compliant	0.991	0.997	0.994	183,844
Non-compliant	0.956	0.892	0.923	15252
accuracy			0.989	199,096
macro avg	0.974	0.944	0.958	199,096
weighted avg	0.988	0.989	0.988	199,096

Generalization is strong, with no substantive evidence of overfitting. The AUC declines only marginally from train to validation (Δ=0.0057) and train to test (Δ=0.0067), while PR–AUC decreases by 0.0112 (train→val) and 0.0375 (train→test), a modest drop consistent with natural distributional shift rather than memorization. Thresholded performance is equally stable: at the standard 0.50 cut, F1 decreases by just 0.011 (train→val) and 0.045 (train→test); at the optimized 0.73 cut, the drops are 0.012 and 0.033, respectively. Three-fold cross-validation further supports robustness, yielding tightly concentrated scores (AUC 0.9804±0.0008, PR–AUC 0.9389±0.0011, F1 0.9071±0.0010) across resamples. Collectively, the small train–test gaps, stability under threshold tuning, and low CV variance confirm that the model's predictive power is not an artifact of overfitting and can be expected to transfer reliably to unseen VAT returns.

Table 7 Overfitting Check

Split	AUC	PR–AUC	F1@0.50	F1@0.73
Train	0.9889	0.9788	0.948	0.942
Validation	0.9832	0.9676	0.937	0.930
Test	0.9822	0.9413	0.903	0.909

The learning curves for both ROC–AUC and Average Precision further confirm the robustness and accuracy of the model. Training scores remain consistently high (ROC–AUC ≈0.996, AP ≈0.98–0.99), while the cross-validation (CV) scores steadily increase with additional training examples, converging toward the training line. The small and stable gap between training and CV curves indicates that the model is not memorizing the training data but instead generalizes effectively to unseen data. This convergence at high performance levels demonstrates strong predictive accuracy and provides further evidence of the absence of overfitting (Figure 3).

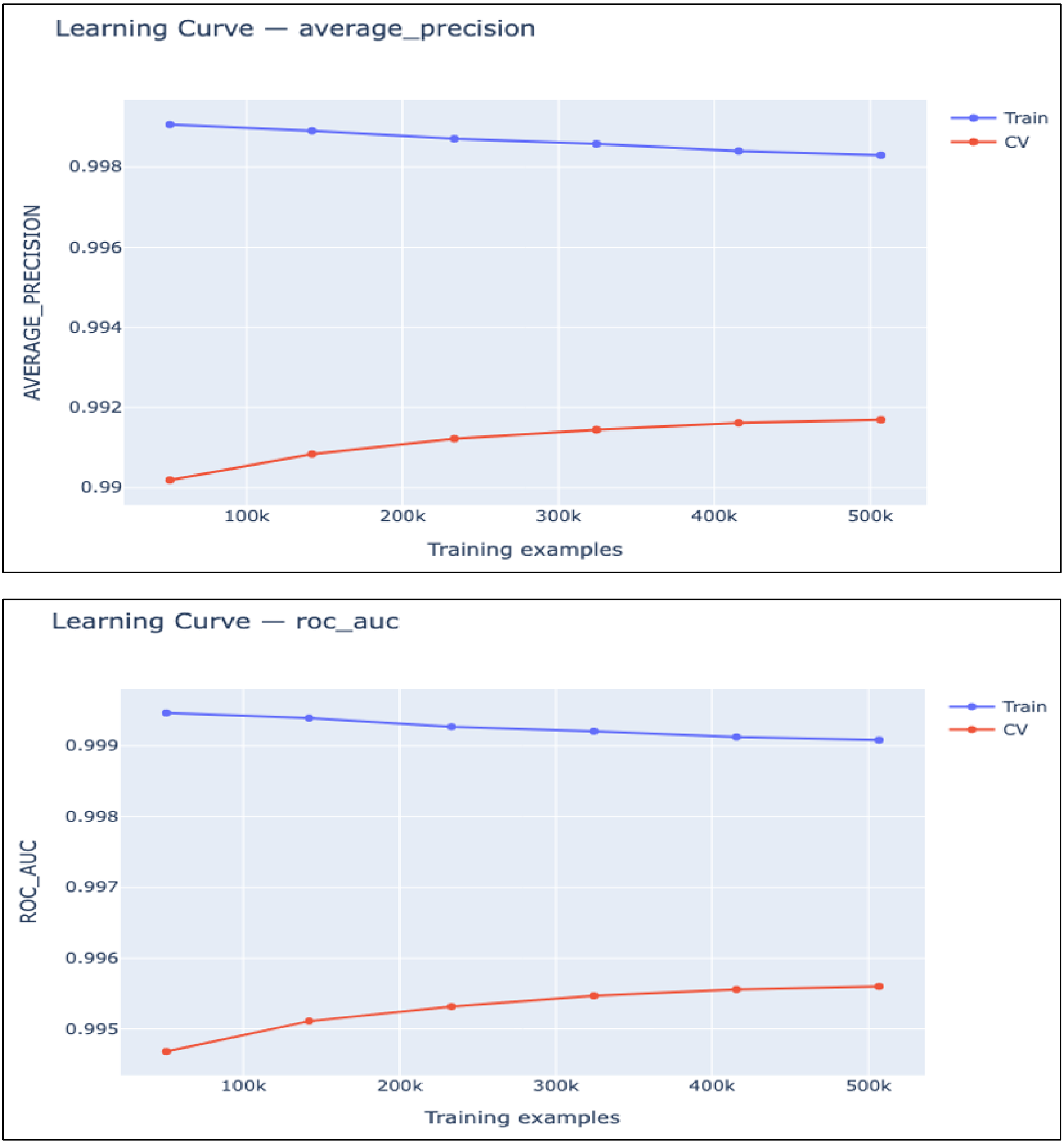

Fig 3 Learning Curve (Left: Average Precision, Right: AUC)

- *Threshold-Sensitive Results and Trade-Offs*

Because the model outputs continuous risk scores, operational performance depends on the decision threshold. Two decision rules are of particular interest for administration: (i) a threshold that maximizes precision subject to a minimum recall constraint; and (ii) a threshold that maximizes the F1-score for the non-compliant class. Under the first rule, the threshold is selected to satisfy a recall floor of 0.85, reflecting the policy preference to capture at least 85% of truly non-compliant cases; among all thresholds meeting this constraint, the one with the highest precision is chosen. Under the second rule, the chosen threshold balances precision and recall to maximize their harmonic mean (F1), providing an efficient compromise when audit capacity is neither extremely scarce nor abundant.

As can be seen from Figure 4, the model accuracy remains consistently high across thresholds, reflecting the model's ability to correctly classify both compliant and non-compliant returns in a highly imbalanced test set. Precision and recall for the non-compliant class follow the expected trade-off, but both remain at high levels within a broad threshold range, which explains the strong F1-score of 0.921. The steep early rise and stabilization of the F1 and accuracy curves indicate that the model quickly reaches an optimal balance and maintains it, suggesting robustness rather than sensitivity to small changes in the decision threshold. In practical terms, this means that the classifier effectively retrieves most non-compliant returns while minimizing false positives among compliant ones, consistent with the high class-wise precision, recall, and macro-averaged F1 reported.

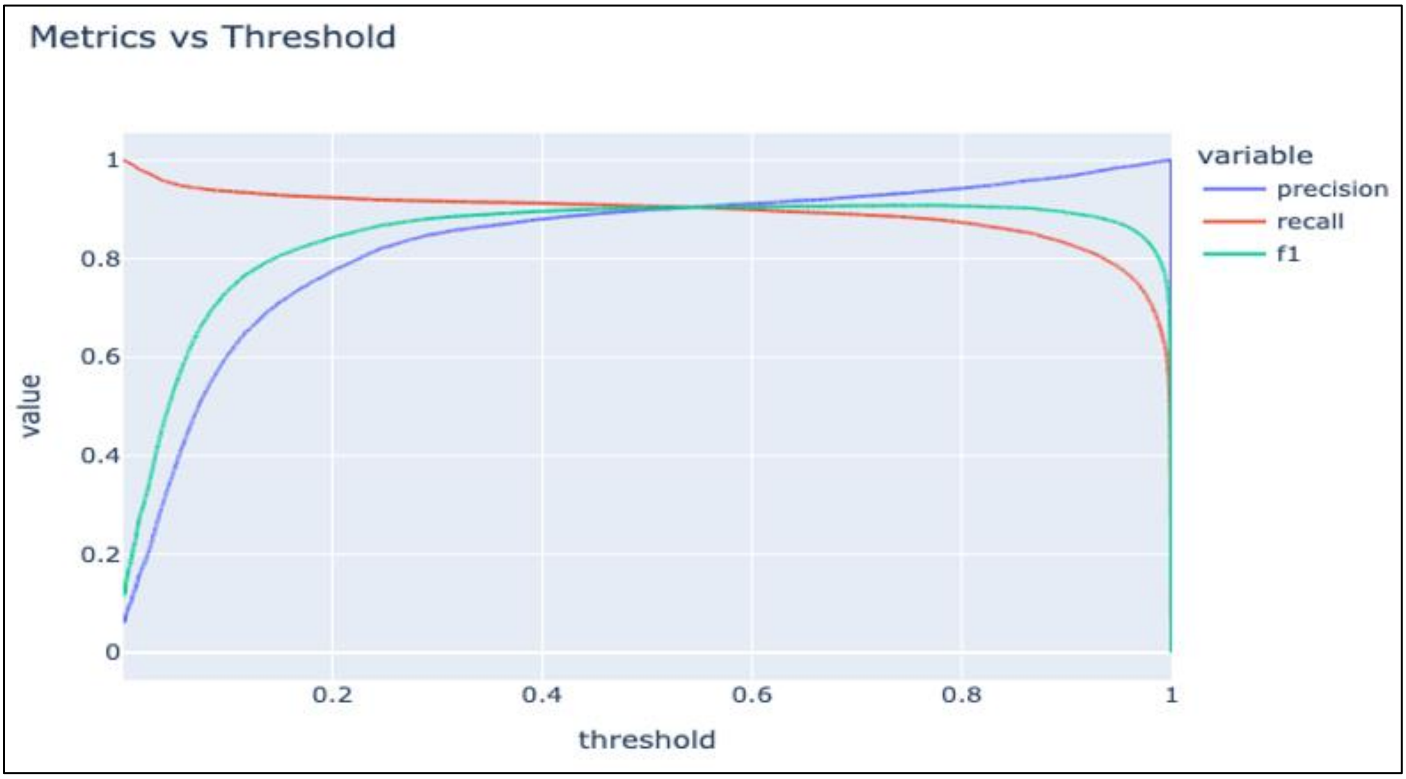

Fig 4 Model Performance Metrics at Different Threshold Point

The empirical results conform to standard detection trade-offs. Relative to a default 0.50 threshold, the recall-constrained threshold slightly increases recall at the cost of a modest rise in false positives, while the F1-optimised threshold yields a balanced profile close to the reported 0.921 F1 for the non-compliant class. Importantly, even under recall-enhancing settings, the model maintains high precision, ensuring that audit resources are not dissipated on large volumes of compliant returns.

- *Error Structure and Practical Interpretation*

The confusion matrix implied by the reported class metrics suggests two small but operationally distinct error types. First, false negatives (non-compliant returns not flagged by the model) constitute roughly 10% of truly non-compliant cases. These are the primary opportunity cost from a deterrence perspective, as undetected cases can persist. Second, false positives (compliant returns erroneously flagged) are well below 1% of compliant filings, limiting audit friction and taxpayer burden. In a risk-based audit design, the very high precision among top-scored cases allows the authority to prioritize the riskiest deciles with confidence, while the small tail of missed cases can be mitigated through random audits and complementary business rules.

- *Feature Importance and Mechanisms*

The feature importance results (Figure 5) show that recorded sales, penalty history and the ratio of turnover reported are the strongest predictors of non-compliance. These variables directly capture discrepancies between what is recorded in the electronic billing system and what is declared in VAT returns, confirming that under-reporting of sales is the primary driver of risk. Other influential features include declared turnover, taxpayer registration and industry codes, and measures of VAT contributions. Lower-ranked variables such as exports value, zero-rated sales, and customs activity volume play a relatively minor role. This ranking reinforces the conclusion that the model’s predictive power is anchored in sales reporting patterns, with EBM-related data providing the clearest signal of VAT non-compliance.

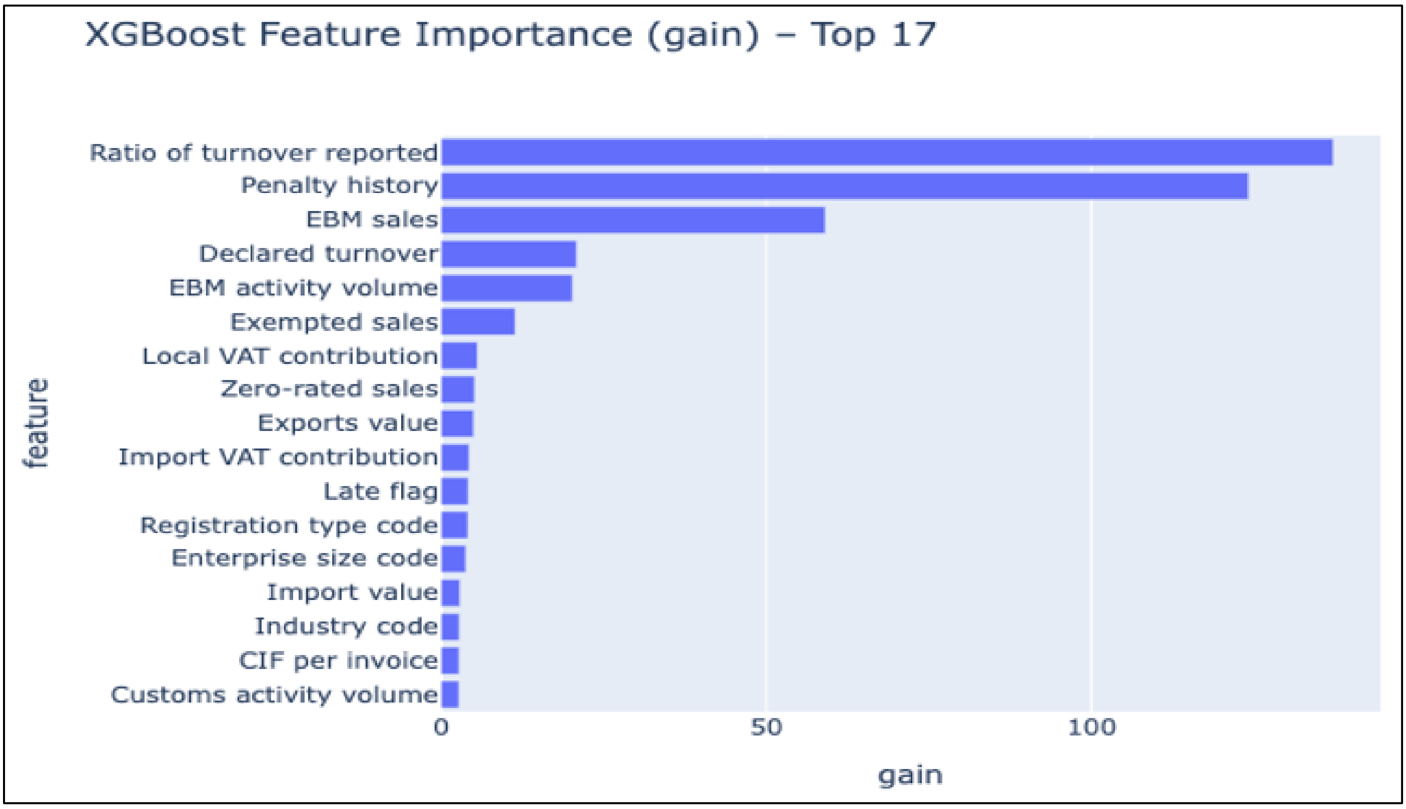

Fig 5 Feature Importance Gain

From an administrative design perspective, these findings validate the integration of third-party data into automated risk scoring. They also suggest where refinements would be most valuable.

F. *Discussion of the Findings*

This study set out to develop an industry-aware machine learning model to predict VAT non-compliance in Rwanda. The findings reveal a nuanced non-compliance landscape, where risk is not uniformly distributed but varies significantly with taxpayer scale and industrial sector. This section dissects three core findings: the counter-intuitive relationship between firm size and VAT non-compliance, the structural drivers of sectoral risk, and the profound predictive power of XGBoost.

➢ *The Paradox of Scale*

This study finds that VAT non-compliance is highest among large taxpayers (11.5%), followed by medium (9.4%) and small taxpayers (6.0%). This directly addresses the second research objective and contradicts the common assumption that Small and Medium Enterprises (SMEs) are the most non-compliant due to informality and resource limitations [16].

The literature explains this pattern through economic complexity. Large firms manage multiple product lines, cross-border transactions, and intricate supply chains [17], which create opportunities for non-compliance in areas such as supply classification, transfer pricing, and input VAT credit claims. These strategies resemble corporate tax avoidance practices documented in prior studies [18].

Deterrence dynamics also differ by scale. For small firms, detection risks impose heavier consequences relative to operations, while for large corporations, potential gains from aggressive tax behavior outweigh risks with penalties often treated as routine costs. The model's finding that penalty history strongly predicts future non-compliance supports this interpretation.

➢ *Sectoral Heterogeneity*

This study finds statistically significant variation in VAT non-compliance across industries, confirming a core research objective. The national average rate of 6.9% conceals elevated risks in Transport and Storage, Wholesale and Retail Trade, and Manufacturing. International evidence similarly identifies these sectors as difficult to administer due to cash-intensive transactions, complex input–output structures, and challenges in tracking goods and services [5].

The model further clarifies sector-specific mechanisms. In Manufacturing, non-compliance stems from multi-stage production and extensive input VAT credits, with errors often tied to classification and EBM mismatches. In Wholesale and Retail, despite the mitigating effect of EBMs on cash-register fraud, misallocation of sales across tax rates persists, partly due to weak consumer incentives prior to the 2024 VAT reward scheme. In Transport and Storage, risks arise from fragmented services, subcontracting, and ancillary fees that are frequently misclassified or omitted.

➢ *The Power and Interpretability of Predictive Analytics*

The model's high performance of 98.9% accuracy and a high F1-score of 0.921 for the VAT non-compliant class affirms the predictive power of machine learning when applied to rich administrative data. This level of accuracy surpasses many benchmarks reported in the literature for financial fraud detection, which often hover in the 85-95% range depending on the context and data quality [15] [19]. The successful rejection of the ho provides robust empirical validation for the utility of this advanced analytical approach in tax administration.

However, predictive accuracy alone is insufficient for policy adoption. The model's true value, and a key contribution of this study, lies in its interpretability, which links directly back to the descriptive findings and policy goals. The feature importance ranking provides a clear, data-driven narrative of VAT non-compliance, including penalty history, and EBM-related features. The high importance of EBM reported sales and the ratio of declared sales to EBM sales validates Rwanda's significant investment in third-party information reporting systems. This aligns with a large body of literature demonstrating that third-party data is one of the most effective tools for improving tax compliance, as it dramatically increases the visibility of economic transactions [20]. The model essentially learns to automate the process of "red flagging" discrepancies that a human auditor would look for but does so systematically and at scale.

➢ *Implications of the Findings*

This study yields several critical implications for both policy and practice. The findings move beyond a mere description of compliance gaps and offer a predictive framework for proactive intervention. The central implication is the ability to shift from a reactive, post-facto enforcement model to a proactive, data-driven one.

The model's ability to accurately predict VAT non-compliance allows RRA to significantly enhance its operational efficiency. Instead of conducting random or routine audits, which consume time and resources with a low success rate, RRA can now prioritize audits based on a risk score. This targeted approach means that auditors and administrative staff can be directed toward the taxpayers and sectors with the highest probability of non-compliance. This will not only increase the audit hit rate but also reduce the administrative burden on compliant businesses, fostering a more positive relationship between taxpayers and the authority. This strategic reallocation of resources can lead to substantial gains in revenue collection without a proportional increase in enforcement costs.

The statistically significant variation in VAT non-compliance across different industries implies that a "one-size-fits-all" approach to tax policy is sub-optimal. RRA can use the model's insights to design and implement tailored compliance strategies for each sector. For instance, industries with a high number of transactions to final consumers like retail or hospitality may require stronger consumer incentives to address the last-mile non-compliance issue. In contrast, sectors characterized by complex supply chains and large business-to-business transactions might benefit more from enhanced cross-verification of invoices and data analytics to detect input-output mismatches. This ensures that interventions are more effective and better aligned with the specific risks of each industry.

On the other hand, the success of the predictive model underscores the immense value of Rwanda's existing digital tax infrastructure, particularly EBM system. The findings demonstrate that the vast amount of data collected by the EBM can be transformed into a powerful tool for policy and enforcement. Building on this, this study serves proves that modern data science techniques are not only applicable but highly effective in the context of developing economies like Rwanda.

IV. SUMMARY, CONCLUSION, AND RECOMMENDATIONS

➢ *Summary of the Study*

This study addresses the persistent challenge of VAT non-compliance in Rwanda by transitioning from a descriptive, retrospective analysis to a predictive, proactive framework. It establishes that while Rwanda has invested in digital tax systems like EBM, significant compliance gaps remain, with under-reporting of sales being a prevalent issue. The research aimed to bridge this gap by developing an industry-aware machine learning model to predict VAT non-compliance. The study's objectives included collecting and pre-processing datasets from RRA, analyzing sectoral compliance variations, and building an XGBoost-based predictive model. The ultimate goal was to provide a tool for risk-based auditing and to offer data-driven policy recommendations. The study successfully demonstrated that predictive analytics can be effectively applied to tax administration in a developing country context, validating the potential of modern data science to augment traditional enforcement methods.

➢ *Conclusion*

This study conclusively proves that a predictive, industry-aware machine learning model can effectively identify and predict VAT non-compliance in Rwanda. The findings confirm the study's central hypothesis that there is a statistically significant variation in VAT non-compliance behavior across different industrial sectors. The developed model achieved an accuracy greater than the hypothesized accuracy of 80%, demonstrating its efficacy as a tool for tax administration. The analysis identified key features that significantly influence non-compliance, EBM sales, reported sales, and others. It founds that VAT non-compliance differs across industry, affirming that a one-size-fits-all approach to tax enforcement is inefficient and that targeted sector-specific strategies are far more effective. The study's results thus provide a powerful, evidence-based mechanism for RRA to move away from reactive enforcement towards a more strategic, data-driven approach to compliance management.

➢ *Recommendations*

Based on the study's findings, the following recommendations are proposed for RRA and policymakers:

- Adopt risk-based auditing: Implement machine learning predictive model to create a risk-scoring system. This will allow RRA to strategically allocate its audit resources, focusing on the taxpayers and sectors identified as high-risk, thereby maximizing revenue recovery and improving the efficiency of enforcement.

- Tailor sector-specific policies: Use the model insights and industry VAT non-compliance benchmark to design customized compliance strategies.
- Invest in data analytics capacity: The success of this model highlights the value of advanced analytics. RRA should invest in building a dedicated data science team or providing training to existing staff to continuously maintain, refine, and expand the predictive models.
- Integrate the model into operations: Design a user-friendly interface that allows auditors and administrators to interact with the predictive model and explore compliance risk insights seamlessly, making it an integral part of their daily workflow.

➢ *Limitations of the Methodology*

Although the study offers significant insights, it is subject to several methodological limitations:

- Data scope: The study focused exclusively on VAT non-compliance related to under-reporting of output VAT. Other forms of non-compliance, such as fraudulent refund claims or income tax evasion, were outside the scope.
- Data availability: The model relied on data from VAT-registered businesses with EBM. The informal sector, sectors not requiring EBM and micro-businesses not registered for VAT were not part of this analysis, which means the model does not capture the full extent of the VAT gap.
- Model black box: Although XGBoostprovides high accuracy, its black box nature makes it challenging to interpret the precise causal relationships behind the predictions, which can sometimes be a barrier to policy trust and adoption.

REFERENCES

[1]. M. Keen, “The nature, importance, and spread of the VAT,” in *The modern VAT*, International Monetary Fund, 2005, pp. 1–22. Available: https://www.elibrary.imf.org/display/book/9781589060265/ch01.xml

[2]. R. De Mooij and A. Swistak, “Value-added tax continues to expand,” *Finance & Development*, vol. 59, no. 1, 2022.

[3]. M. Keen, “Taxation and development — again,” *IMF WORKING PAPERS*, 2016.

[4]. D. Ntihemuka, C. Niyomugabo, J. C. Nshimiyimana, U. I. Grace, F. Hakizimana, and C. Harushyubuzima, “TAX STATISTICS IN RWANDA: FISCAL YEAR 2023/24 - 7th edition,” Rwanda Revenue Authority, 2024.

[5]. L. Ebrill, M. Keen, J.-P. Bodin, and V. Summers, Eds., *The modern VAT*. International Monetary Fund, 2001.

[6]. T. Siwale, B. Dillon, J. Mulenga, and K. Musole, “Harnessing the power of electronic fiscal devices to increase VAT revenue in zambia: The role of consumers and consumer incentive schemes,” International Growth Centre (IGC), Policy Brief ZMB-20014, 2021.

[7]. E. Ghirmai, S. Logan, and S. MuRRAy, “The incidence and impact of electronic billing machines for VAT in rwanda.” International Growth Centre, 2016.

[8]. G. Mascagni, D. Mukama, and F. Santoro, “An analysis of discrepancies in taxpayers’ VAT declarations in rwanda,” 2019.

[9]. P. F. Mugambwa and O. Habineza, “Tax alert: Rwanda moves to make electronic invoicing system (EIS) invoices mandatory for all businesses!” PwC, 2021.

[10]. N. Hakizimana and F. Santoro, “Technology evolution and tax compliance: Evidence from rwanda,” workingpaper, 2024.

[11]. O. Tuyishimire and B. F. Murorunkwere, “Applications of big data analytics in tax compliance monitoring: A case study of rwanda’s value-added tax,” *CESifo Economic Studies*, 2024.

[12]. C. MUNEZERO, “Value added tax fraud detection using naïve bayes data mining approach,” Master's Thesis, University of Rwanda, College of Business; Economics, Kigali, Rwanda, 2020.

[13]. M. Battaglini, L. Guiso, C. Lacava, D. L. Miller, and E. Patacchini, “Refining public policies with machine learning: The case of tax auditing,” National Bureau of Economic Research, Cambridge, MA, Working Paper Series 30777, 2022. Available: http://www.nber.org/papers/w30777

[14]. South African Revenue Service (SARS), “ SARS uncovers non-compliance through data-driven risk detection.”

[15]. A. Kamoun, R. Boujelbane, and S. Boujelben, “A prediction model to detect non-compliant taxpayers using a supervised machine learning approach: Evidence from tunisia,” *Journal of Business Analytics*, vol. 8, no. 2, pp. 116–133, 2025, Available: https://doi.org/10.1080/2573234X.2024.2438195

[16]. M. A. Umar, C. Derashid, I. Ibrahim, and Z. Bidin, “Public governance quality and tax compliance behavior in developing countries,” *International Journal of Social Economics*, vol. 46, no. 3, pp. 338–351, Oct. 2018, doi: 10.1108/IJSE-11-2016-0338.

[17]. K. P. Chen and C. Y. C. Chu, “Internal control versus external manipulation: A model of corporate income tax evasion,” *The RAND Journal of Economics*, vol. 36, no. 1, pp. 151–164, 2005.

[18]. M. Hanlon and S. Heitzman, “A review of tax research,” *Journal of Accounting and Economics*, vol. 50, no. 2–3, pp. 127–178, 2010.

[19]. J. West and M. Bhattacharya, “Intelligent financial fraud detection: A comprehensive review,” *Computers & Security*, vol. 57, pp. 47–66, 2016.

[20]. D. Pomeranz, “No taxation without information: Deterrence and self-enforcement in the value added tax,” *American Economic Review*, vol. 105, no. 8, pp. 2539–2569, 2015.