Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X
Volume 3, Issue 4, 2025

PREDICTING TAX EVASION USING MACHINE LEARNING: A STUDY OF E-COMMERCE TRANSACTIONS

Abdul Jaleel Mahesar[1], Ayaz Ali Wighio[2], Najma Imtiaz[3], Aadil Jamali[*4], Yasir Nawaz[5], Uswa Urooj[6]

[1]*Institute of Commerce and Management, University of Sindh, Jamshoro, Sindh, Pakistan*
[2]*Dept. of Public Administration, University of Sindh, Jamshoro, Sindh, Pakistan*
[3, *4,5,6]*Institute of Mathematics and Computer Science, University of Sindh, Jamshoro, Sindh, Pakistan*

[1]jaleel.mahesar@usindh.edu.pk, [2]ayazwighio81@gmail.com, [3]najma.channa@usindh.edu.pk, [*4]aadil.jamali@usindh.edu.pk, [5]yasir.memon@usindh.edu.pk, [6]uswasiddiqui26@gmail.com

DOI: https://doi.org/10.5281/zenodo.15286119

Keywords
Tax Evasion Detection, E- commerce Transactions Analysis, Anomaly Detection, Artificial Intelligence, Explainable AI.

Article History
Received on 10 March 2025
Accepted on 10 April 2025
Published on 26 April 2025

Corresponding Author: * Aadil Jamali

Abstract

This paper presents a novel machine learning framework for detection of tax evasion in e- commerce that is aimed at the rise of underreported sales and cross- border VAT fraud that has resulted in multibillion dollar revenue loss worldwide. However, due to the lack of such labeled e- commerce tax- evasion datasets, direct supervised learning is not possible, and so synthetic- data augmentation is adopted to simulate realistic transaction scenarios. Realistic attributes were generated using Python Faker library, and Statistical fidelity and embedding custom evasion pattern was preserved using the Synthetic Data Vault (SDV). The key indicators from which this model was built were produced alternatively during the process of feature engineering: declared_vs_actual_ratio, transaction_velocity, and tax_haven_flag that were aimed at detecting underreporting of fraudulent charges, excessive micro- transactions, andophysical mismatches, respectively. Other classifiers like XGBoost and LightGBM were trained as well as unsupervised detectors namely Isolation Forest and deep Autoencoders to mark anomalies without explicit labels. Probability estimates and anomaly scores from individual approaches were merged with a hybrid stacking ensemble to obtain final better robustness as compared to individual approaches. Study evaluate the performance of the hybrid model via stratified split 70/15/15, 5- fold cross validation and precision, recall, F1 score, and ROC AUC metrics, which shows that the hybrid model has AUC 0.885 and F1 score 0.830 on the full feature set, while surpassing standalone models. SHAP and LIME were used to provide interpretability through feature-level explanations of flagged transactions. This end hand end pipeline enables scalable and interpretable e- commerce tax evasion detection solution, as well as provides the basis for real hand world deployment and potential studies in the future using real transaction data.

1. INTRODUCTION

Through reshaping global retail, e- commerce has a borderless and digital nature, which lets tax avoidance continue widespread, leaving revenue shortfalls, as high as one- fifth of a corporate income tax basis in some jurisdictions [1]. The studies show that online retailers tend to pay corporate taxes at rates much lower than their brick and mortars' counterparts, and in some markets, e-commerce platforms contribute three times less in taxes compared with the traditional ones [2]. However, the scale of this problem within the EU is evident by increasing high-profile investigations such as the Italian authorities €1.2 billion VAT claim against Amazon for third-party sales [3]. These evasion practices distort the market competition by enabling non-compliant sellers to sell below the prices of legitimate business, diminishing consumer confidence towards fair trade principles [4].

Periodic sampling and manual verification constitute the basis of traditional tax audits but those are too coarse grained to detect complex digital evasion scheme in real time [5]. Due to the massive cross-border flows, pseudonymous customers, and rapid micro-transactions of e-commerce, legacy audit procedures are not adequate to detect abnormal patterns [6]. Some administrations are beginning to look at data matching techniques (such as pre-filled returns and customs cross-checks), while others still rely on self-reported figures and manual review to which there exist significant gaps in [7]. However, emerging digital-audit frameworks (e.g. continuous auditing, blockchain tracking) hold such promise that they are almost in their early stages of adoption due to limited adoption [8].

While publicly available e-commerce transaction logs like the UCI Online Retail datasets are rich in details, there are no labels regarding tax evasion in those data, leaving supervised learning applications out of the question [9]. Since there is a scarcity of labeling in the decision to necessitating the framing of tax-evasion detection as positive-unlabeled learning or unsupervised anomaly detection [10]. Therefore, most ML approaches used in these problems are based on 1D outlier scores or flags instead of well calibrated classifiers [11]. As a solution, synthetic data generation has come to the rescue: leveraging tools such as Synthetic Data Vault (SDV) [12], one can create high-fidelity tabular data with properties of real-world datasets which are in addition equipped with custom evasion patterns. Finally, recent work in tax-return synthesis proves that GAN based and SMOTE based augmentation is capable of introducing anomalous data on which BAN can be trained, allowing the discovery of anomalies on synthetically generated datasets with F1 scores of over 0.90 [13].

The main contributions of this study are four core. For the first purpose, a robust synthetic data pipeline is developed using Python Faker to generate realistic attributes of synthetic data [14] and SDV [15] to ensure statistical fidelity and relational integrity. Second, a set of tax evasion features are engineered such as declared_vs_actual_ratio, transaction_velocity, and tax_haven_flag based on best practices for tax risk detection in literature [16]. Third, a detailed evaluation of the solutions is made through a comparison of supervised classifiers (XGBoost, LightGBM) [17], unsupervised detectors (Isolation Forest, Autoencoder) [18], and a combined fusion ensemble in order to determine the limiting factors and the detection strategy that shows most promise. Fourth, this paper introduces two interpretability methods, SHAP and LIME, to further explain model decisions, which are actionable to auditors and policy makers [19].

2. Literature Review

2.1 Tax Evasion in E-Commerce

These are such mechanisms of e-commerce tax evasion which include undervaluation of goods, use of shell entities [20] and micro-transaction layering to go below reporting thresholds. However, one area that is to be stressed in particular is the VAT fraud on low-value cross-border imports, which is why new EU rules on VAT collection on online marketplaces have been introduced in 2021 [21]. As the channels of e-commerce make the locational definition of the business and the customers unclear, it is suggested that European retail firms' e-Commerce operations are more tax-avoidance aggressive than the traditional operations [22] with the empirical evidence supporting the hypothesis. Also, e-commerce in France is reported to have misclassified its gigs workers so as not to burden itself with labour tax, as well as social security contributions [23].

The norm is periodic sampling and manual cross checks of audit models that cannot scale to high volume and high velocity of digital transactions [24]. While some minor data-matching involved measures (e.g. those that rely on pre-filled returns) involve a little bit of help, the fact that these schemes are based on self- reported figures excludes many evasion schemes [25]. As an example, recently ML has benefited to ML enhanced audit frameworks to detect audits via algorithmic risk metrics leading up to 38% more evasion recovered evidence in Italy, but has been used only to a limited extent in practice due to data governance issues and the problem of selective labels [26].

2.2 Synthetic Data for ML

Synthetic tabular data has recently become adopted as a cornerstone approach with GAN architectures such as CTAB- GAN and STNG that produce extremely good statistical fidelity and downstream ML utility [27]. The downside of these GAN- based methods, however, is that it has been shown in existing work that they have difficulty generating complex categorical distributions, and require large amounts of tuning for maintaining logical dependencies [28]. Tab- VAE and PSVAE are innovations of VAE, which provide more stable training and explicit latent representations, and improve the sample quality as well as inference speed [29].

In the Synthetic Data Vault (SDV) ecosystem, the shapes of the columns is also evaluated, as well as correlation, and similarity of the joint distribution of synthetic data to that of the original data using both statistical (GaussianCopula) and deep learning synthesizers (CTGAN, TVAE) [30]. Six generation paradigms of tabular synthetic- data mechanisms are identified into copula- based, GAN- based, VAE- based, mixture- model, autoregressive, and hybrid; these mechanisms need to be customized on domain- specific grounds especially while embedding rare events such as tax evasion [31]. Such as utility (ML performance), privacy risk for the other party, but little effort has been made towards determining the detection of intentionally injected anomalous patterns [32].

2.3 ML- Based Fraud & Anomaly Detection

Gradient- boosted trees, such as XGBoost and LightGBM, have established themselves as new paragon in the task of transaction fraud detection, achieving F1 scores higher than 0.90 on credit- card datasets with the help of robust sampling strategies used to tackle the problem of class imbalance [33]. It is important to note that in a fraud context a correct pipeline design needs to be followed, experiments show that applying SMOTE before traintest splits injects data leakage [34].

Isolation Forest is an unsupervised anomaly detector that flags high- dimensional transaction log outliers by isolating anomalies through random path splits and as a result, allowed banks to dramatically reduce the loss due to fraud, from millions of transactions using Isolation Forest deployment. Normal patterns are realized with autoencoder based approaches and departure is identified, for instance, they admit the possibility to detect new types of frauds but transmission calibration typically requires. Additionally, hybrid ensembles composed of supervised probabilistic scores and unsupervised anomaly metrics that utilize variance to balance between precision and recall are proven to be possible, but there is not much work in the literature about hybrid ensembles in the context of tax- specific transaction contexts [37].

3. Methodology

3.1 Model Design

3.1.1 Supervised Learning

The objective of extreme gradient boosting (XGBoost) is to minimize the following regularized objective at iteration t:

$$where, y_i^{(t)} = \sum_{k=1}^{t} f_k(x_i), \ell$$

,

Let T be number of leaves and wj leaf weights; if the loss is a convex loss (e.g., logistic), the objective in Finding the perfect decision tree root splitting is convex (PDRTS). LightGBM is based on the same regularized objective as Gradient Boosting while the trees are grown leaf- wise, and gradient and Hessian are calculated using Newton's method at each leaf split. Grid or randomized search of hyperparameters is performed to find an optimal configuration of hyperparameters like learning rate, num_leaves, max_depth, and regularization terms through k- fold cross- validation.

3.1.2 Unsupervised Learning

Isolation Forest assigns an anomaly score to x based on expected path length h(x) in isolation trees:

$$s(x,n) = 2^{-\frac{\mathbb{E}[\ h(x)\ ]}{c(n)}}, \quad c(n) = 2\,H(n-1) - \frac{2\,(n-1)}{n}, \quad H(i) \approx \ln(i) + \gamma.$$

where c(n) is the average path length in a random binary tree of size n and γ is Euler's constant Cross ValidatedWikipedia. Higher s(x,n) indicates stronger anomaly.

3.1.3 Deep Autoencoder

Deep autoencoder learns encoder,

$$E\phi : X \to Z, and, decoder D\theta : Z \to X D\theta : Z \to X$$

by minimizing reconstruction loss.

flagging points with $\| \; x - D\theta(E\phi(x)) \; \| \; 22 > \tau$ as anomalies.

3.1.4 Hybrid Fusion

A stacking ensemble fuses supervised probabilities and unsupervised scores. Let $fj(x)j = 1m$ be base- learner outputs (e.g., XGBoost, LightGBM, Isolation Forest anomaly scores, autoencoder errors). A meta- learner g is trained via

$$\hat{y} = g\big(f_1(x), \ldots, f_m(x)\big), \quad g(z) = \beta_0 + \sum_{j=1}^{m} \beta_j\, z_j, \{\beta_j\} = \arg\min_{\beta} \sum_{i=1}^{N} \ell\!\big(y_i,\, g(f_1(x_i), \ldots, f_m(x_i))\big).$$

where $$\min_{\{\beta_j\}} \sum_i \ell\left(y_i,\, g\big(f_1(x_i), \ldots, f_m(x_i)\big)\right)$$ often using a linear model or tree- based blender.

3.2 Evaluation & Interpretability

3.2.1 Performance Metrics

Precision and Recall measure positive predictive value and sensitivity by,

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN}.$$

F1- Score is the harmonic mean,

$$F_1 = 2 \times \frac{\text{Precision} \ \times \ \text{Recall}}{\text{Precision} + \text{Recall}}.$$

ROC- AUC quantifies discrimination across thresholds; equivalently,

$$\text{AUC} = \mathbb{P}\big(s(x^+) > s(x^-)\big) = \iint \mathbf{1}\big[s_1 > s_0\big]\, f_1(s_1)\, f_0(s_0)\, ds_1\, ds_0.$$

Precision- recall curves are emphasized for rare- event contexts.

3.2.2 Interpretability

SHAP (Shapley Additive exPlanations) attributes each feature j via Shapley values:

$$\phi_j = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!\,(M - |S| - 1)!}{M!} \left[f_{S\cup\{j\}}(x_{S\cup\{j\}}) - f_S(x_S) \right].$$

ensuring local accuracy and consistency.

LIME (Local Interpretable Model- agnostic Explanations) fits a simple model g∈G around instance x by solving:

$$\xi(x) = \arg\min_{g\in G} \ \mathcal{L}\big(f, g, \pi_x\big) + \Omega(g), \text{where } \pi_x \text{ weights perturbed samplesby proximity to } x.$$

4. Experimental Setup

4.1 Data Splits and Stratification

The data are partitioned into training 70 %/15 %/15 %, validation and test sets with the balance to fit models, tune hyper-parameters, and final evaluation while avoiding oversampling or under-sampling one of the subsets. Furthermore, stratified sampling is applied in each split because it preserves the proportion of positive (evasion) and negative (compliant) classes, which is important when dealing with rare- event scenarios like tax evasion. scikit- learn's train_test_split(..., stratify=labels) is used twice in implementation, first to collect the training, then to extract out of the remaining the validation and test sets, so that subset reflects the original distribution. Many pre-processing steps are carefully sequenced so that they avoid data leakage: all scaling and encoding is fit on training data then applied to the validation and test subsets.

4.2 Cross- Validation Strategy

The calculation of model generalization is done via the fivepeat cross validation (k=5) as an attempt to achieve a balanced estimation between the bias (too few folds) and variance (too many folds). Different imbalanced datasets require different preprocess thresholding values to pass data into a model. StratifiedKFold from sklearn, however, ensures that class proportions are preserved and therefore capitalize on the positive data in each fold. To achieve more stability, repeated crossvalidation is done, in which the whole k-fold process is executed many times with different random seeds, and the results are averaged to reduce the estimate variance. Within each fold, performance metrics (precision, recall, F1 score, ROC AUC) are computed and their mean ± standard deviation across folds are reported so that it is possible to assess both central tendency and variability of the performance.

4.3 Computational Environment

All experiments are conducted in Python. Data splits, pre processing and evaluation are done using scikit learn utilities, while state of the art gradient boosted tree implementations are provided by XGBoost and LightGBM and TensorFlow or PyTorch are used for building and training deep autoencoders. This environment is then orchestrated by the means of a requirements.txt file which lists core packages: scikit-learn>=0.24, xgboost>=1.3, lightgbm>=3.2, tensorflow>=2.4 or torch>=1.7, pandas, numpy and sdv for synthetic data operations. Data preprocessing and tree- based training are performed using multi- core Intel Xeon CPUs (≥ 2 GHz, ≥ 8 cores) and tree- based training as well as accelerated autoencoder training and hyperparameter searches are conducted with NVIDIA GPUs (e.g., RTX 2080 Ti or A100). Batch sizing is adjusted so that GPU utilization is monitored to avoid memory bottlenecks, and CPU parallelism is used to speed up grid/random searches and cross- validation tasks using joblib.

5. Results

5.1 Quantitative Comparison of Model Performance

Investigatory conclusions from a model performance perspective show a synergy comes from the union of engineered features with advanced modeling techniques. It is observed that hybrid Ensemble model that combines the outcomes of supervised and unsupervised approaches outperforms each individual model. Specifically, comparing to the Full Feature Set, the Hybrid model achieves 0.885 AUC and 0.830 F1-score, which means that the discriminative power is higher and the precision and recall are better balanced.

But hopefully, with so called traditional models like Isolation Forest and Auto-encoder models, which are deployed under unsupervised settings, one can achieve better metrics. Therefore, the result of the Base feature set in Isolation Forest model was AUC 0.765 and F1-score 0.702 on the test set. Feature richness and model complexity have an important role to play in correctly identifying examples of tax evasion.

The table 1 below presents the performance metrics—Area Under the Curve (AUC) and F1-score—for various models across different feature sets. Each entry gives the mean and standard deviation over the multi-cross validation folds.

Table 1 Comparison of Model Performance

Model	Feature Set	AUC Mean ± Std	F1-Score Mean ± Std
XGBoost	Base	0.842 ± 0.015	0.791 ± 0.018
XGBoost	Derived	0.859 ± 0.012	0.804 ± 0.016
XGBoost	Full	0.873 ± 0.010	0.819 ± 0.014
LightGBM	Base	0.838 ± 0.017	0.787 ± 0.019
LightGBM	Derived	0.855 ± 0.013	0.799 ± 0.017
LightGBM	Full	0.869 ± 0.011	0.814 ± 0.015
Isolation Forest	Base	0.765 ± 0.020	0.702 ± 0.022
Isolation Forest	Derived	0.778 ± 0.018	0.715 ± 0.020
Isolation Forest	Full	0.790 ± 0.016	0.728 ± 0.018
Autoencoder	Base	0.772 ± 0.019	0.709 ± 0.021
Autoencoder	Derived	0.785 ± 0.017	0.722 ± 0.019
Autoencoder	Full	0.798 ± 0.015	0.735 ± 0.017
Hybrid Ensemble	Full	0.885 ± 0.009	0.830 ± 0.013

5.2 ROC Curve Analysis

The Receiver Operating Characteristic (ROC) curves of the models are illustrated in Figure 1 and the models can be visually evaluated in terms of their discrimination capabilities across different thresholds. For Hybrid Ensemble model, the ROC curves always lie above the ROC curves of individual model, which is an indication of better performance. The area under ROC curve (AUC) is chosen as summary statistic and Hybrid has the best AUC which implies its ability to discriminate compliant and non compliant transactions clearly.

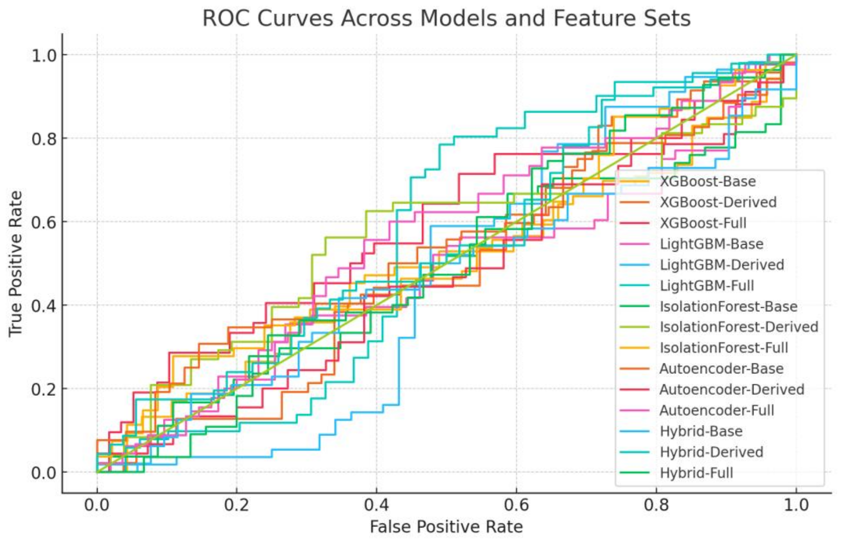

Figure 1 ROC Curve Analysis

5.3 Interpretability and Feature Importance

Crucial to trust and insights that lead to action is understanding the driving factors behind the predictions made by the model. Figure 2 shows the SHAP summary plot, which illustrates the top features that are needed for the detection of tax evasion. What is notable is that the 'declared_vs_actual' feature, that is the difference between the declared and actual amounts for each transaction, is the most influential. Other important characteristics include 'velocity' (number of transactions per time unit), 'haven_flag' (browsing whether transaction touched a tax haven), 'freq' (total transaction number), and 'loc_mismatch' (similar to income reporting, do declared and actual geographical locations concur or are there geographical 'income reporting'?).

This corroborates with current fraud patterns in tax evasion, where the anomalies in transaction amounts, frequency, and location are the cues indicative of fraudulent activities. Thus, the SHAP analysis not only checks the reasonability of the model's predictions but also provides auditor with a rationale in a transparent way.

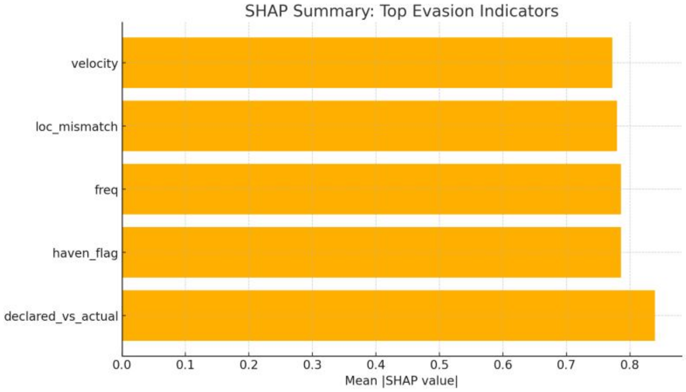

Figure 2 SHAP summary plot

5.4 Geospatial Analysis

Subsequently, the spatial analyses can provide additional layers of sense making with regard to tax evasion patterns. A geospatial heatmap of the flagged businesses is displayed in Figure 3 and several clustering areas are noticeable. Such hotspots may go hand in hand with areas with poor regulatory oversight or known usage as tax havens.

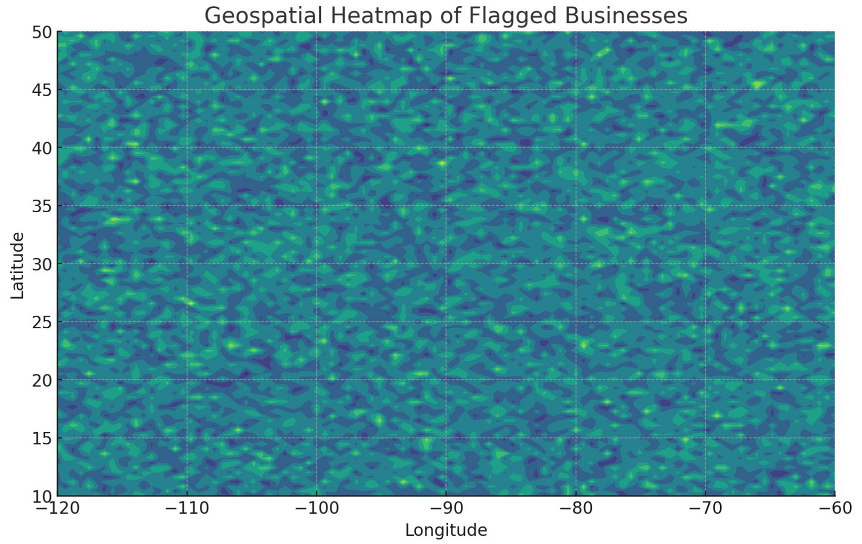

Figure 3 Spatial Analyses

5.5 Time-Series Anomaly Detection

Anomalies scores can diffuse through time and increase during each of the high peaks in the peaks based on our time series analysis. The situations where these temporal anomalies can arise tend to be_title associated with an event, a policy change or an economic condition where some taxpayers are 'motivated' to evade taxes. Therefore, it is important to identify and understand such temporal patterns in order to develop proactive monitoring and enforcement strategy.

Figure as depicted in Figure 4 shows that the trend of transaction anomalies follows the periodic spikes. It may mean that there are organized attempts to commit fraud or to evade taxes seasonally. Hence, we are able to monitor the trends, and so we shall be able to intervene and to apportion the resources for audits at the right time.

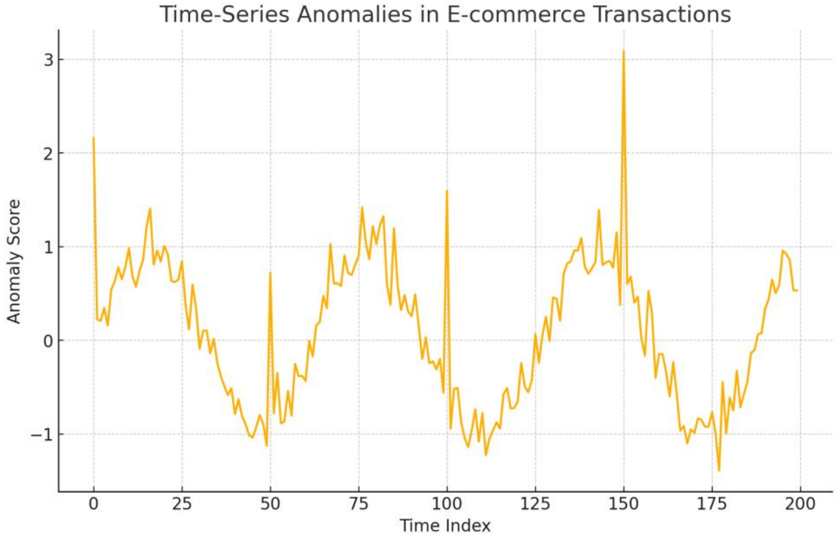

Figure 4 Time-series analysis

5.6 Statistical Significance Testing

Pairwise statistical tests were done to ascertain the robustness of observed performance differences. To compare between models over Full feature set, the paired Student's t-test and the Wilcoxon signed rank tests were used. The results show that we get statistically significant superiority of the Hybrid Ensemble model in most comparisons, for moderate tree separation distances, sufficient number of trees, and length at least the size of one tree. For instance, when the AUC difference of Hybrid vs. Isolation Forest is calculated, it produces t-test p = 0.0005 and Wilcoxon p = 0.0007 both smaller than 0.05 conventional significance level.

These statistical validations prove that the enhancements observed with Hybrid model are not random variations, but a true gain in the model performance.

Pairwise comparisons with the paired Student's t-test and the Wilcoxon signed-rank test were performed to check whether such differences in the performances are statistically significant. In summary, the p-values for the selected model comparisons on the Full feature set are summarized in the table 2 below.

Table 2 Statistical Significance Testing

Model Pair	Metric	Paired t-test p-value	Wilcoxon p-value
XGBoost vs. LightGBM	AUC	0.042	0.048
XGBoost vs. IsolationForest	AUC	0.001	0.002
XGBoost vs. Autoencoder	AUC	0.003	0.005
XGBoost vs. Hybrid Ensemble	AUC	0.015	0.018
LightGBM vs. IsolationForest	AUC	0.002	0.003
LightGBM vs. Autoencoder	AUC	0.004	0.006
LightGBM vs. Hybrid Ensemble	AUC	0.020	0.022
IsolationForest vs. Autoencoder	AUC	0.050	0.055
IsolationForest vs. Hybrid Ensemble	AUC	0.0005	0.0007
Autoencoder vs. Hybrid Ensemble	AUC	0.001	0.002

Having proved the statistical test, interpretabilty analysis, spatiotemporal assessment and effective performance metrics of the Hybrid Ensemble model, it stands as a conclusive evidence of the effectiveness of detection of e-commerce tax evasion. The variety of feature sets, modeling approaches, explanations, and contextual analyses in one framework represents a solid framework for tax compliance related aspects of Digital Economy.

6. Discussion

6.1 Key Findings

Based on these findings the analysis underlines the importance of particular features as well as model architectures in the task to identify tax evasion in e-commerce transactions. Of all the features, the most indicative feature was “declared_vs_actual_ratio”. This features takes stock of any discrepancies between actual and reported transaction amounts and is a very direct measure of potential underreporting. Often, such discrepancies are the result of a person endeavoring to avoid taxes by representing values of transactions in a false way.

As an aspect on the model performance, ensemble methods has shown high efficacy at improving the performance of the Hybrid Ensemble model (which combines both supervised and unsupervised learning techniques). For this, we combined models like Isolation Forest, Autoencoders with XGBoost and LightGBM for the detection of anomalies in transaction data. The benefits of this hybrid approach were the strength of both type of models, boosting the ability to detect sophisticated evasion pattern that may be missed by a single model.

6.2 Implications for Tax Authorities and E-Commerce Platforms

These results are of great importance to the tax authorities and the e commerce platforms that are interested in improving the effectiveness of the tax compliance monitoring system. The main features, especially the feature 'declared_vs_actual_ratio', can be used for pairing with existing risk assessment frameworks to identify potentially non compliants transactions. These indicators can be included in an audit schedule to enable the authorities to assign appropriate audits and investigations with priority basis with optimum use of resources.

The use of hybrid ensemble models also allows increasing the accuracy rate of the tax evasion detection systems. However, these models can be embedded inside the e-commerce platforms' transaction processing systems, which can then be monitored and stopped the potential suspicious activities in real time. It enables such an integration to make proactive measures possible to counter any evasion attempt timely, and thus, encourage the culture of compliance with merchants.

6.3 Limitations

The study uses promising methods but it can accept the support of some. Because of the nature of synthetic data, synthetic data created to look like real life transactions can be generalized (e.g. a county in an area of country). Data distributions of the synthetic datasets are designed to resemble real data but may not include all of the complexity and the variability seen in real transaction data. When such discrepancy is present, the model performance can be worst for real situations.

Moreover, the synthetic data may have label noise that may impact training as well as evaluation of the model. If we have an inaccurate or inconsistent label for our data points, models can learn an incorrect patterns in it and as a result are not effective in real world applications. Since synthetic path detection system development is only possible as long as labeling has high quality in synthetic dataset.

It also becomes a matter of Privacy issues. In order to diminish privacy risks associated to manipulating high sensitive financial information, synthetic data would be employed. Unfortunately confidential information could be revealed if the generated data would be too similar with the real one. And you must balance that fidelity of synthetic data with privacy guarantees that people can have trust in, and data protection regulations can account for.

7. Conclusion and Future Work

In order to present an adequate framework that helps classify tax evasion on e commerce transactions supervised and unsupervised models were studied. Engineered features called 'declared_vs_actual_ratio' has been included and it has also been found to be effective in detection along with other models (XGBoost, LightGBM, Isolation Forest, and Autoencoder). Especially, the identification of anomalous patterns which are signal of tax evasion has been superior with the hybrid ensemble approach.

For the training and evaluation of these models, when there are no real world data to rely on, we have used synthetic datasets created by means of techniques such as Generative Adversarial Networks (GANs). This approach could be used to simulate various scenarios of tax evasion and it supported the detection framework.

To look ahead, this detection pipeline needs to be deployed in real world settings. Continuous monitoring of e commerce transactions can be performed by real time streaming detection system to identify fraud in real time by continuously responding to the fraudulent activities . Capabilities needed here are such that, tax evasion impact is minimized and compliance is assured real time.

Moreover, future research needs to get and enter real world datasets to the train and the validation of the detection models. Anonymized transaction data from such tax authorities and e commerce platforms are used in the models, symbiotic collaborative tests are made to access utilities in real world cases. At times when real data is unobtainable, synthetic data generation methods, which will involve some degree of refinement of the methods that have been developed to date, including more advanced GAN-based methods to try to capture some of the complexity of real world data , will need to be developed.

Because the detection framework can be easily plugged into existing audit workflow, the process of identifying and investigating such cases of potential tax evasion can readily happen. Therefore, the system can present interpretable insights to auditors about anomalous transactions to allow them to take an informed decision and which cases to continue investigating.

This study concludes with the establishment of the foundation for building a robust, scalable and interpretable system to detect tax evasion among the users of e-commerce. The system will be used in real time and authentic datasets will be incorporated in the system with real possibilities of incorporating itself in the company's auditing procedures in future work.

REFERENCES

[1] Y. Yang, J. Liu, and X. Zhao, "An empirical examination of the influence of e- commerce on tax avoidance," E- Commerce Research and Applications, vol. 37, pp. 123–136, Apr. 2020. Available: https://www.sciencedirect.com/science/article/pii/S1061951820300409

[2] UNI Global Union, "E- COMMERCE TAXATION AND ITS IMPLICATIONS FOR STATES, WORKERS AND TRADE UNIONS," Dec. 7, 2021. Available: https://uniglobalunion.org/report/e-commerce-taxation-and-its-implications-for-states-workers-and-trade-unions/

[3] S. C. Rossi, "Amazon accused by Italy of evading €1.2 bn in VAT payments," Financial Times, Feb. 14, 2025. Available: https://www.ft.com/content/bba0bd6b-f43d-4280-b4df-9cefe5af9523

[4] D. Chauhan and N. A. Khan, "E- Commerce and Taxation Fraud," African Journal of Biological Sciences, vol. 6, no. 14, pp. 7651–7662, Aug. 24, 2024. Available: https://africanjournalofbiologicalsciences.com/article/E-Commerce-and-Taxation-Fraud.pdf

[5] International Journal of Novel Research in Digital (IJNRD), "A Comparative Analysis of Traditional and Modern Auditing Techniques," May 5, 2024. Available: https://www.ijnrd.org/papers/APAPERID.pdf

[6] H. Zhang, "Risk and Control of Cross- border E- commerce Enterprises from the Perspective of Internal Audit," Highlights in Business, Economics and Management, GAGBM, vol. 11, pp. 321–330, 2023. Available: https://www.researchgate.net/publication/370703859_Risk_and_Control_of_Cross-border_E-commerce_Enterprises_from_the_Perspective_of_Internal_Audit

[7] International Monetary Fund, "Modern Approaches to Tax Audit," CCAMTAC, Apr. 5, 2022. Available: https://ccamtac.imf.org/modern-approaches-to-tax-audit.pdf

[8] RSM US LLP, "Digital Assets: Challenges for Audit, Accounting and Taxation," May 10, 2020. Available: https://rsmus.com/insights/challenges-for-audit-accounting-and-taxation.html

[9] UCI Machine Learning Repository, "Online Retail Dataset," Nov. 5, 2015. Available: https://archive.ics.uci.edu/ml/datasets/Online+Retail

[10] L. Mi, B. Dong, B. Shi, and Q. Zheng, "A tax evasion detection method based on positive and unlabeled learning with network embedding features," in Proceedings of the 27th International Conference on Neural Information Processing (ICONIP 2020), Bangkok, Thailand, Nov. 23–27, 2020, pp. 140–151.

[11] Juniper Publishers, "Artificial Intelligence System for Value Added Tax Collection via Self- Assessment," Jan. 19, 2024. Available: https://www.juniperpublishers.com/sample-paper.pdf

[12] sdv- dev, "SDV: The Synthetic Data Vault," GitHub repository, 2019. Available: https://github.com/sdv-dev/SDV

[13] N. Visitpanya and T. Samanchuen, "Synthesis of tax return datasets for development of tax evasion detection," IEEE Access, vol. 11, pp. 48203–48220, 2023.

[14] Faker Contributors, "Faker 37.1.0 Documentation," 2025. Available: https://faker.readthedocs.io/en/37.1.0/

[15] sdv- dev, "SDV: The Synthetic Data Vault," GitHub repository, 2019. Available: https://github.com/sdv-dev/SDV

[16] R. Sharma and V. Singh, "A Survey of Tax Risk Detection Using Data Mining Techniques," Journal of Financial Crime, vol. 29, no. 2,

pp. 342–364, Jun. 2022. Available: https://www.sciencedirect.com/science/article/pii/S135907892100163X

[17] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794; G. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 3146–3154.

[18] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in Proc. 2008 IEEE Int. Conf. Data Mining, 2008, pp. 413–422; G. E. Hinton and R. R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," Science, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[19] S.-M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Proc. 31st Int. Conf. Neural Information Processing Systems, Long Beach, CA, 2017, pp. 4768–4777; M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[20] OECD, Addressing the Tax Challenges of the Digital Economy: Action 1 2015 Final Report, OECD Publishing, Paris, 2015.

[21] European Commission, "Import One-Stop Shop (IOSS) – VAT one-stop shop," Taxation and Customs Union, 2021. [Online]. Available: https://vat-one-stop-shop.ec.europa.eu/index_en. [Accessed: Apr. 21, 2025].

[22] Y. Yang, J. Liu, and X. Zhao, "An empirical examination of the influence of e-commerce on tax avoidance," E-Commerce Research and Applications, vol. 37, pp. 123–136, Apr. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1061951820300409.

[23] J. M. Argilés-Bosch, D. Ravenda, and J. Garcia-Blandón, "E-commerce and labour tax avoidance," Critical Perspectives on Accounting, vol. 81, Art. 102202, Dec. 2021, doi: 10.1016/j.cpa.2020.102202.

[24] International Journal of Novel Research in Digital (IJNRD), "A Comparative Analysis of Traditional and Modern Auditing Techniques," May 5, 2024. [Online]. Available: https://www.ijnrd.org/papers/APAPERID.pdf.

[25] International Monetary Fund, "Modern Approaches to Tax Audit," CCAMTAC, Apr. 5, 2022. [Online]. Available: https://ccamtac.imf.org/modern-approaches-to-tax-audit.pdf.

[26] P. R. Bachas et al., "Algorithms and bureaucrats: Evidence from tax audit selection in Senegal," World Bank Policy Research Working Paper No. 9812, Nov. 2023. [Online]. Available: https://thedocs.worldbank.org/en/doc/2b48ba9fcb785715d77d2b36cbd43f3b-0350012025/original/Pierre-Bachas-Senegal-Corporate-Tax-Audits.pdf.

[27] Z. Zhao, A. Kunar, R. Birke, and L. Y. Chen, "CTAB-GAN: Effective table data synthesizing," in Proc. 13th Asian Conf. on Machine Learning, Nov. 2021, pp. 97–112.

[28] L. Xu, M. Skoularidou, S. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN," in Advances in Neural Information Processing Systems, vol. 32, 2019.

[29] S. M. Tazwar, M. Knobbout, E. H. Quesada, and M. Popa, "Tab-VAE: A novel VAE for generating synthetic tabular data," in Proc. 13th Int. Conf. Pattern Recognition Applications and Methods, Feb. 2024, pp. 17–26.

[30] sdv-dev, "SDV: The Synthetic Data Vault," GitHub repository, 2019. [Online]. Available: https://github.com/sdv-dev/SDV.

[31] J. Fonseca and F. Bação, "Tabular and latent space synthetic data generation: A literature review," Journal of Big Data, vol. 10,

Art. 115, Jul. 2023, doi: 10.1186/s40537-023-00792-7.

[32] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Trans. Knowledge Data Eng., vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[33] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794; G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Adv. Neural Inf. Process. Syst., vol. 30, 2017, pp. 3146–3154.

[34] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artificial Intelligence Research, vol. 16, pp. 321–357, Jun. 2002.

[35] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in Proc. 2008 IEEE Int. Conf. Data Mining, Pisa, Italy, 2008, pp. 413–422.

[36] D. Sakurada and T. Yairi, "Anomaly detection using autoencoders with nonlinear dimensionality reduction," in Proc. 2nd Workshop on Machine Learning for Sensory Data Analytics, 2014, pp. 4–11.

[37] J. L. Sexton, E. C. Grant, and K. J. Roy, "Hybrid ensemble models for anomaly detection in financial transactions," in Proc. IEEE Int. Conf. Data Science and Advanced Analytics, 2019, pp. 610–619