JURNAL RISET INFORMATIKA **P-ISSN: 2656-1743 |E-ISSN: 2656-1735**
Vol. 8, No. 4. September 2026 DOI: https://doi.org/10.34288/jri.v8i4.651
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

XGBOOST-BASED CLASSIFICATION OF POTENTIALLY FRAUDULENT SHOPEE SELLERS USING SELLER METADATA AND SHAP

Azka Fahsya[-1], Wargijono Utomo [-2], Harry Dwiyana Kartika[-3]

Teknik Informatika / Fakultas Teknik
Universitas Krisnadwipayana
azkafahsya2104@gmail.com[-1], wargiono@unkris.ac.id[-2], harry_d@unkris.ac.id[-3]

Abstract

The rapid growth of e-commerce increases the exposure to fraudulent sellers, while reactive reporting mechanisms generally operate only after consumers suffer losses. This study develops a proactive fraud-risk screening model to assist in identifying potential fraud-risk indications among Shopee sellers from store-profile metadata and to explain the model's decisions. Primary data were collected from Indonesian Shopee seller profiles using a manual-hybrid approach with Playwright. The initial dataset contained 460 profiles and was reduced to 457 clean records after duplicate removal. The final dataset comprised 355 Safe stores and 102 Fraud stores (labeled heuristically using external indicators) with 11 predictor features. XGBoost was trained with class weighting through a scale_pos_weight value of 3.48 and an 80:20 stratified split. Evaluation on 92 test records produced 95.65% Accuracy, 84.00% Precision, 100.00% Recall, and 91.30% F1-Score. SHAP analysis identified historical_sold (48.42%), store age (18.73%), and rating_star (14.76%) as the three most influential features. The trained model was subsequently integrated into a Chrome Extension prototype to provide predictions and SHAP-based rationales. The results indicate that a seller-centric metadata approach can support early fraud-risk detection; however, the 84.00% precision confirms that the output should be interpreted as a risk screening indication rather than an absolute determination of fraud, assisting consumers in making safer decisions.

Keywords: *Fraud Detection*, *E-Commerce*, *Shopee*, *Sellers*, *XGBoost*, *Explainable AI*, *SHAP*

Abstrak

Pertumbuhan pesat e-commerce meningkatkan paparan terhadap penjual yang melakukan penipuan, sementara mekanisme pelaporan reaktif umumnya baru beroperasi setelah konsumen mengalami kerugian. Penelitian ini mengembangkan model penapisan risiko penipuan (*fraud-risk screening*) proaktif untuk membantu mengidentifikasi potensi indikasi risiko penipuan pada penjual Shopee berdasarkan metadata profil toko serta menjelaskan keputusan model. Data primer dikumpulkan dari profil penjual Shopee Indonesia menggunakan pendekatan manual-hibrida dengan Playwright. Dataset awal terdiri dari 460 profil dan dikurangi menjadi 457 data bersih setelah penghapusan duplikasi. Dataset akhir terdiri dari 355 toko berlabel Safe dan 102 toko berlabel Fraud (dilabeli secara heuristik menggunakan indikator eksternal) dengan 11 fitur prediktor. XGBoost dilatih dengan pembobotan kelas menggunakan nilai scale_pos_weight sebesar 3,48 dan pembagian data terstratifikasi 80:20. Evaluasi terhadap 92 data pengujian menghasilkan Accuracy sebesar 95,65%, Precision sebesar 84,00%, Recall sebesar 100,00%, dan F1-Score sebesar 91,30%. Analisis SHAP mengidentifikasi historical_sold (48,42%), umur toko (18,73%), dan rating_star (14,76%) sebagai tiga fitur yang paling berpengaruh terhadap prediksi model. Model yang telah dilatih kemudian diintegrasikan ke dalam prototipe Chrome Extension untuk memberikan prediksi dan penjelasan berbasis SHAP. Hasil penelitian menunjukkan bahwa pendekatan berbasis metadata penjual dapat mendukung deteksi dini risiko penipuan; namun, nilai Precision sebesar 84,00% menegaskan bahwa keluaran model harus diinterpretasikan sebagai indikasi penapisan risiko, bukan sebagai penentuan mutlak bahwa suatu penjual melakukan penipuan, sehingga dapat membantu konsumen membuat keputusan yang lebih aman.

Kata kunci: Deteksi Penipuan, E-Commerce, Penjual, Shopee, XGBoost, Explainable AI, SHAP

The work is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License

INTRODUCTION

Digital transformation has shifted trade activities from physical interactions to platform-based transactions. E-commerce expands market access, but the lack of direct interaction also creates room for business identity manipulation. On Shopee, consumer protection is hindered by reactive fraud handling mechanisms that heavily rely on victim reports, giving perpetrators the opportunity to erase their digital footprints (Awalia et al., 2024). Responding to the increase in losses due to fraud, Karim (2025) emphasizes the urgency of a computational approach to identify risks before material losses occur. Technically, monitoring should not be limited to transaction history; the operational metadata of the seller's profile (store age, sales history, and rating) is crucial for distinguishing organic stores from fictitious ones.

The integration of AI in e-commerce has been widely explored to mitigate risks and protect consumers from evolving digital threats (Hertati & Yadiati, 2024; Kustanto, 2024). Furthermore, advanced analytics such as graph neural networks and victim-centric frameworks have recently been proposed to enhance fraudulent-seller detection and analyze their future trends (Adams, 2024; Farsi & Chowdhury, 2025). Previous fraud research on marketplaces has proven the effectiveness of machine learning in learning anomaly patterns. (Ardhitha et al., 2025; Dekou et al., 2022) used ensemble algorithms to detect fraud based on user behavior variables and digital transactions. Yin and Li (2025) and Miruna et al. (2026) also validated the superiority of XGBoost for e-commerce fraud detection. Although proven effective, the majority of these studies still focus on buyer behavior and the use of secondary datasets, rather than on seller identity (seller-centric).

In addition to the unit of analysis issue, fraud data generally has a skewed (imbalanced) class distribution. Handling imbalanced data is a critical challenge in fraud detection, often requiring specific algorithmic treatments or ensemble deep learning strategies to prevent bias (Lhaura et al., 2024; Mohammed et al., 2025). XGBoost was chosen because its tree-based architecture has proven robust for tabular data (Grinsztajn et al., 2022), combined with the scale_pos_weight parameter to maintain detection sensitivity in the Fraud class. Empirical analyses have shown that ensemble methods like XGBoost and Random Forest offer an optimal bias-variance tradeoff for e-commerce datasets (Ranglani, 2024; Salsabila & Hutagalung, 2025) .

Given that XGBoost's decision complexity is a black-box, the Explainable AI (SHAP) approach is integrated to improve transparency. Comparative studies highlight SHAP as a robust method for interpreting models compared to other techniques like LIME (Raufi et al., 2024; Kaya et al., 2026). Incorporating this explainability bridges the gap between complex model outputs and user understanding (Rahim & Basheer, 2025), allowing feature contributions to be read globally and locally so that classification decisions can be accounted for (Awosika et al., 2023; Thanathamathee et al., 2024). The primary data extraction process in this study used a manual-hybrid method with Playwright. Human intervention is applied at the authentication stage, while browser automation directly extracts metadata to bypass dynamic rendering and the platform's anti-bot mechanisms (Chen, 2024; Stafeev & Pellegrino, 2024).

Previous studies on e-commerce fraud detection have predominantly focused on transaction behavior, buyer-side activity, or secondary datasets. Limited attention has been given to seller-centric classification based on publicly observable store profile metadata combined with explainable machine learning. Therefore, this study addresses this gap by proposing an XGBoost-based seller fraud-risk screening framework integrated with SHAP to provide both predictive results and interpretable explanations.

Based on the problem formulation and the urgency described above, this research aims to achieve the following measurable objectives:

1. To develop an XGBoost model for classifying seller fraud-risk based on store profile metadata.
2. To evaluate model performance using Accuracy, Precision, Recall, and F1-Score under imbalanced conditions.
3. To identify the contribution of specific seller metadata features globally and locally using SHAP.
4. To implement the finalized model as a functional Chrome Extension prototype for client-side risk screening.

RESEARCH METHODS

Research Type and Design

This research is an experimental quantitative study using the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, ensuring a comprehensive approach from initial data acquisition to final deployment

The work is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License

(Shimaoka et al., 2024) . Computational experiments were executed using Python in a Jupyter Notebook, and the final model was deployed as a Chrome Extension prototype. While deploying machine learning models via browser extensions is an effective client-side security measure against fraud (Wu et al., 2023) , ensuring their discoverability remains a practical challenge (Bucci & Li, 2023) .

Architecturally, the flow in Figure 1 ensures that the computational engineering aligns with business objectives, producing an output that is both mathematically accurate and functionally deployable. Computational experiments were executed using Python (v3.x), XGBoost (v3.2.0), and SHAP (v0.51.0). The XGBoost hyperparameter configuration retained default settings (max_depth=6, learning_rate=0.3, eval_metric='logloss'), combined with class penalty compensation (scale_pos_weight) to handle data imbalance.

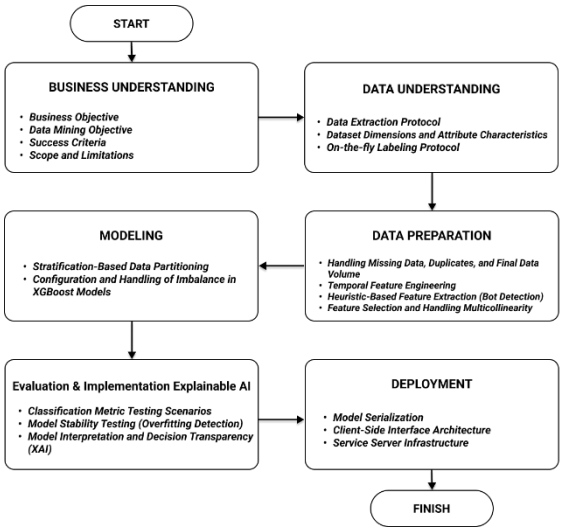


Figure 1. Research flow and system development stages.

Data Collection

The data used is primary data directly extracted from the Shopee Indonesia ecosystem (shopee.co.id) during the March to April 2026 period. Acquisition was performed using a manual-hybrid approach utilizing Chrome Remote Debugging and the Playwright library. The manual intervention stage was utilized to perform standard user authentication and navigate public-facing interfaces to ensure compliance with platform access policies. Once the profile page reached a stable state, Playwright automation was used strictly to read publicly available metadata, maintaining ethical research boundaries without retrieving any private consumer information (Chen, 2024; Stafeev & Pellegrino, 2024).

Data was collected using a purposive sampling approach guided by heuristic searches on social media (X/Threads). The seller selection was conducted broadly across various product categories and geographic regions in Indonesia without being restricted to a specific niche. The inclusion criteria required the stores to be actively accessible during the collection period with fully observable operational metadata. Conversely, stores that were already suspended or hidden by the platform prior to data extraction were excluded. The initial 460 profiles were filtered for duplication to obtain a clean population of 457 records. These profiles serve as an experimental research sample and do not represent the true statistical distribution of all Shopee Indonesia sellers.

Table 1. Research dataset characteristics.

Component	Value
Initial data	460 store profiles
Cleaned data	457 store profiles
Safe class	355
Fraud class	102
Initial feature columns	18
Final feature columns	11
Data split	80% train : 20% test
Training set	365
Testing set	92
Random state	42

Based on the details in Table 1, the final data matrix leaves 457 store profile observations with 11 predictor features columns. It is clear that the target class distribution shows an imbalance (imbalanced dataset), where the Safe class (355 observations) dominates the Fraud class (102 observations).

Population and Sample

The research population comprises all seller profiles registered on the Shopee Indonesia platform. The research sample was determined using a purposive quota sampling method to ensure balanced representation across the platform's diverse market. Specifically, the Shopee Indonesia ecosystem features 25 main product categories. To achieve a comprehensive distribution, approximately 18 to 19 store profiles were extracted from each of these 25 categories, yielding the initial targeted sample of 460 profiles.

The work is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License

The inclusion criteria required these selected stores to be actively accessible with fully observable operational metadata and a broad geographic distribution across Indonesia. Conversely, stores that were already suspended, hidden, or had incomplete public metadata were excluded.

It is explicitly acknowledged that the final validated sample of 457 sellers does not represent the true statistical distribution of the entire Shopee Indonesia seller population. However, the sample size and categorical distribution are highly justified for this experimental modeling phase. Empirical research by Grinsztajn et al. (2022) confirms that gradient-boosted decision trees (GBDT) like XGBoost possess high sample efficiency on tabular data, enabling them to converge optimally on small-to-medium datasets (such as our ~450 instances) without requiring tens of thousands of records. Considering the research focus is seller fraud, the unit of analysis is exclusively limited to store profile metadata and eliminates other variables such as text reviews, product images, or transaction behavior from the buyer's side.

Labeling Technique

The target status is_fraud is not explicitly available from the platform's public API endpoints; therefore, a heuristic on-the-fly labeling protocol was implemented during data acquisition. To minimize subjectivity, labeling was guided by specific structural criteria rather than random assumptions. The "Safe" (Class 0) label was assigned strictly to store entities demonstrating organic operational indicators, such as a consistent historical sales record, active authentic buyer reviews, or official platform verification status (e.g., Star Seller or Shopee Mall).

Conversely, the "Fraud" (Class 1) label was assigned based on external cross-validation and specific anomaly thresholds. A store was classified as Fraud only if it met multiple concurrent criteria: documented consumer complaints on external social media (X/Threads) specifically calling out the store's identity, an overwhelming dominance of negative ratings without resolved disputes, and structural characteristics typical of bot accounts (e.g., highly randomized alphanumeric usernames paired with zero historical sales or extremely premature store age).

While this structured heuristic protocol reduces random assignment, it must be acknowledged that this manual labeling process relies on single-annotator community consensus. The absence of multi-annotator cross-validation and formal inter-rater reliability metrics (such as Cohen's Kappa) introduces potential subjective label noise. Consequently, the labels represent high-confidence heuristic risk indications rather than legally verified fraud data, which stands as a primary methodological limitation of this study.

Data Preparation

The preparation stage focused on cleaning and adjusting the dataset dimensions to be ideal for modeling. Attributes acting as unique identifiers such as url, shop_name, and shop_id were strictly eliminated from the predictor matrix because they lack statistical generalization power. The original username attribute was not used directly as a predictor; instead, prior to the train-test split, it was transformed into a numerical username_score feature using Shannon Entropy, and the original text attribute was subsequently dropped. The tanggal_dibuat attribute was also not maintained in string format, but was prepared for a time transformation process.

The extraction of insights from commercial data and rigorous exploratory data analysis (EDA) are essential for identifying hidden patterns before modeling (Maresti et al., 2024; Tanasă et al., 2023; Wijaya, 2025). Furthermore, effective feature engineering and data preprocessing using Python are fundamental principles for optimizing machine learning models in artificial intelligence applications (Teoh & Rong, 2022; Zheng & Casari, 2024).

Feature Engineering

Feature engineering was applied to extract latent information from raw metadata into two new predictor variables. First, the tanggal_dibuat attribute was converted into umur_toko (store age) in days to represent operational duration maturity. Second, the textual username attribute was converted into a numerical username_score using Shannon Entropy theory to quantify the randomness level of the character arrangement forming the account name. The entropy calculation considers the probability distribution of all alphanumeric characters within the username string.

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i \qquad (1)$$

The higher the resulting entropy value, the more random the character arrangement, which heuristically indicates automated bot account

The work is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i4.651
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

creation. The empirical justification for this assumption and its actual predictive usefulness are subsequently validated through the model's SHAP feature contribution analysis.

Feature Selection

Feature selection was conducted to prevent mathematical ambiguity and interpretability issues due to multicollinearity among independent variables. Although tree-based models like XGBoost are inherently robust to multicollinearity, removing highly correlated features simplifies the model and ensures stable SHAP interpretations. This redundancy test was executed by mapping the Pearson correlation matrix, utilizing a correlation threshold of > 0.80 to flag highly redundant variables.

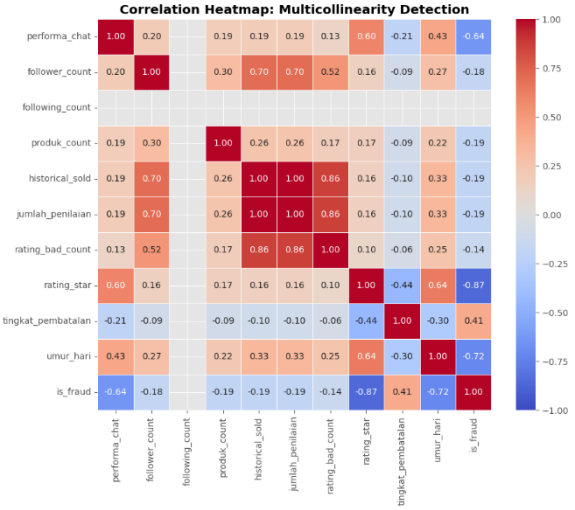


Figure 2. Pearson correlation heatmap for feature redundancy identification.

Based on the correlation matrix visualization in Figure 2, a perfect positive correlation (1.00) was found between the jumlah_penilaian and historical_sold features, as well as a very strong correlation (0.86) involving rating_bad_count. In addition, the following_count feature has a constant value of zero across all instances. To maintain decision tree stability, these three features were eliminated. The historical_sold feature was intentionally retained over jumlah_penilaian because actual sales volume serves as a more direct, factual economic indicator of a store's operational scale than user reviews.

XGBoost Modeling

Modeling was executed using Python's xgboost library (v3.2.0) and scikit-learn (v1.8.0). To prevent overfitting and computational complexity, default hyperparameters were intentionally retained (max_depth=6, learning_rate=0.3, eval_metric='logloss'). The clean dataset was partitioned into 365 training data instances and 92 test data instances (80:20 ratio). The stratify=y and random_state=42 parameters were applied to maintain target class distribution proportions consistently and maintain experimental reproducibility. To overcome class imbalance, mathematical penalty compensation was injected via the scale_pos_weight parameter calculated based on the ratio of majority to minority instances, resulting in a weight value of 3.48.

$$scale_pos_weight = \frac{N_{negative}}{N_{positive}} \quad (2)$$

The algorithm learning process operates by minimizing a loss function integrated with a regularization penalty to suppress tree complexity so the model avoids overfitting.

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \quad (3)$$

Model Evaluation

Model performance testing was executed exclusively on the test dataset. The measurement parameters are based on the Confusion Matrix which is calculated into four main metrics: Accuracy, Precision, Recall, and F1-Score.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

Considering the asymmetric (imbalanced) nature of the dataset, the Accuracy metric is only used as a performance baseline. The critical evaluation is deliberately focused on Precision, Recall, and the F1-Score. In fraud detection contexts, prioritizing Recall ensures the system minimizes detection failures (False Negatives) on the Fraud class, which carries a higher business cost. While additional threshold-independent

The work is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License

metrics such as ROC-AUC, PR-AUC, or repeated cross-validation could theoretically provide broader generalization bounds for small datasets, the current evaluation metrics (Recall and F1-Score) were strictly selected as they directly measure the prototype's primary functional objective: acting as a highly sensitive initial risk-screening tool.

Explainable AI with SHAP

The SHAP method was implemented to open the black-box nature of the XGBoost algorithm. Interpretation was done globally to map the dominant variable hierarchy, and locally to justify detection reasons on one specific entity. The computational foundation of this method rests on Shapley values.

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} [f_x(S \cup \{i\} \quad (8)$$

Global feature contributions are calculated using the Mean Absolute SHAP Value, formulated mathematically as follows:

$$\frac{1}{N} \sum_{i=1}^{N} |\,\phi_{i,j}| \quad (9)$$

where $\phi_{i,j}$ represents the Shapley value of feature j for instance i. This metric quantifies the average magnitude of a feature's impact on model predictions. To facilitate user comprehension, these absolute SHAP values were normalized into percentage distributions. Furthermore, while SHAP effectively explains both the magnitude and direction of feature effects on the model's predictions, these values must be strictly interpreted as correlational feature importance, avoiding any causal interpretations of fraudulent behavior. The extraction of these scores is then converted into a natural language text representation to be understood by users.

Prototype Implementation

As the research downstream, the XGBoost modeling object and SHAP TreeExplainer were serialized and embedded on a cloud server (Hugging Face Spaces) using the Flask API framework. This system is controlled from the client side via a Chrome Extension, tasked with capturing store profile metadata in real-time, sending it to the server, and rendering the final result in the form of security status, probability level, and top 3 detection reasons (Top 3 Features) on the consumer's browser screen. As a proof-of-concept demonstrator, the current system successfully establishes the real-time classification pipeline. Comprehensive software system evaluations including API prediction latency, cross-browser compatibility, extensive usability testing, and robustness to continuous changes in the Shopee webpage structure are acknowledged as technical limitations that will be addressed in future development iterations.

RESULTS AND DISCUSSION

Dataset Characteristics

The initial dataset consisted of 460 store profiles. A comprehensive missing-value analysis confirmed that there were no null values across all feature columns. Regarding outlier analysis, since the classification utilizes a tree-based ensemble algorithm (XGBoost) that is inherently robust to extreme values, outliers in continuous features (such as highly inflated historical_sold numbers) were intentionally retained to preserve the natural representation of hyper-active or anomalous stores. Descriptive evaluations of feature distributions highlighted distinct structural differences: Safe stores generally exhibited significantly higher operational maturity (umur_toko) and consistent transaction volumes, whereas stores in the Fraud class frequently displayed premature operational ages paired with anomalous or non-existent transaction histories.

Data Cleaning Results

After duplication inspection, three identical instances were removed, yielding a final matrix of 457 clean profiles ready for modeling. The data cleaning process produced a highly consistent final matrix. The removal of identity-based attributes avoided the use of variables functioning merely as entity recognizers. Furthermore, the redundancy evaluation revealed a perfect positive correlation (1.00) between jumlah_penilaian and historical_sold, and a constant zero variance in following_count. Eliminating these redundant features effectively reduced mathematical ambiguity and information duplication.

Class Distribution

Understanding the distribution of target classes is a crucial step in evaluating the dataset's structural integrity before proceeding to the modeling phase. After completing the data cleaning process, the final data matrix distribution consists

The work is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i4.651
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

of 355 Safe stores and 102 Fraud stores. The visual representation of this disparity between the two categories is clearly illustrated in Figure 3.

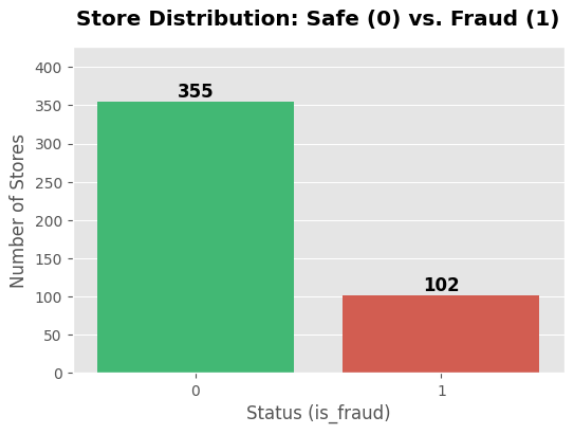


Figure 3. Class distribution in the cleaned dataset: 355 Safe and 102 Fraud.

As shown in Figure 3, this proportion highlights a distinct class imbalance, which necessitates specialized computational handling. This asymmetric distribution makes the model potentially biased toward the majority class if all misclassification errors are penalized equally. Therefore, the application of the scale_pos_weight parameter set to 3.48 is explicitly directed to increase the algorithm's sensitivity toward the minority Fraud class. This approach is consistent with the research objectives, prioritizing Recall and F1-Score as the primary performance indicators over global Accuracy.

Feature Engineering Results

The transformation of tanggal_dibuat into umur_toko changes temporal information into numerical variables that the model can learn. Meanwhile, the username_score transforms textual username attributes into a quantifiable entropy measure. Statistical evaluation of this engineered feature confirms its empirical usefulness; observations indicate that the average username_score for the Fraud class is characteristically higher compared to the Safe class, validating the hypothesis that bot-generated or fictitious accounts frequently utilize highly randomized alphanumeric sequences. Thus, two pieces of information that were originally not directly compatible with a numerical matrix were successfully converted into valid quantitative predictors.

Table 2. Feature Engineering Results

Instance	Raw Attribute (Before Transformation)	Engineered Result (After Transformation)
Store 1	tanggal_dibuat: 2019-11-20 *username*: "guming123"	umur_toko: 2379 hari *username*_score: 2.95
Store 2	tanggal_dibuat: 2020-04-27 *username*: "cahayalistrikid"	umur_toko: 2220 hari *username*_score: 3.27

Feature Engineering Results

The redundancy evaluation successfully identified features with high multicollinearity. Based on the Pearson correlation matrix, the jumlah_penilaian and rating_bad_count features were eliminated due to their strong positive correlation with historical_sold (1.00 and 0.86, respectively). Additionally, the following_count feature was dropped as it contained constant zero values across all instances, offering no predictive variance. This strict selection process resulted in a final, mathematically optimized matrix of 11 independent predictor features.

XGBoost Configuration

Training was conducted on 365 training instances and 92 testing instances using a random_state of 42. Logloss monitoring indicated that the test data error dropped significantly from 0.47150 in the initial iterations to 0.10113 around the 30th iteration, before stabilizing at 0.10186 by the 99th iteration. Concurrently, the training data logloss at the final iteration was recorded at 0.00369. This learning curve pattern demonstrates a stable convergence in the conducted experiment, confirming that the model successfully mapped the classification boundaries without experiencing divergence.

Model Evaluation Results

Test results on 92 store profiles showed Accuracy reached 95.65%, but this measure is not used as the sole basis of success because the dataset is imbalanced. Precision of 84.00% means not all Fraud predictions are actual Fraud; in other words, false positives still exist. Conversely, a Recall of 100.00% indicates that the model successfully captured all Fraud instances explicitly within this

The work is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License

specific 92-instance test set. An F1-Score of 91.30% shows a high balance between Precision and Recall.

Table 3. Results of the XGBoost model evaluation on the test data.

Metric	Result	Interpretation
Accuracy	95.65%	Overall classification accuracy on test data
Precision	84.00%	Proportion of fraud predictions that were actual fraud cases
Recall	100.00%	The model successfully detected all Fraud instances within the test set
F1-*Score*	91.30%	Balance between Precision and Recall

To dissect specifically where these prediction errors lie, evaluation is mapped using a confusion matrix.

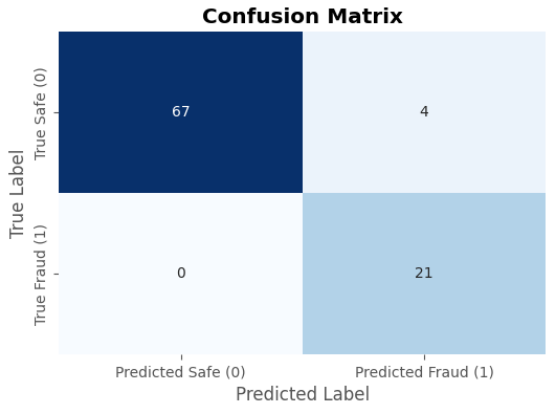


Figure 4. Confusion matrix of prediction results for 92 test data points.

The confusion matrix shows no False Negatives in the Fraud class, aligning with the 100.00% Recall. However, the 84.00% Precision (4 False Positives) indicates that a strategy highly sensitive to the Fraud class still produces some wrong positive predictions. Operationally, the consequence is that the system is more suitable as an initial risk screening tool rather than as a final determination mechanism for a store's status.

Model Stability Analysis

Training data accuracy was recorded at 100.00%, while test data was 95.65%, resulting in a performance gap of 4.35%

```
=== OVERFITTING EVALUATION ===
Training Data Accuracy : 100.00%
Testing Data Accuracy  : 95.65%
Performance Gap        : 4.35%

Model indicates stable fit within this split.
```

Figure 5. Comparison of accuracy performance for training data and test data.

Based on empirical best practice heuristics (Ranglani, 2024), this gap is below the 5.00% threshold, which indicates stability within this specific 80:20 data split. However, achieving a 100% training accuracy suggests the model may be memorizing the training set, meaning this gap alone is insufficient to definitively claim the model is entirely free from overfitting. This evaluation remains limited to a single split, so the stability and robustness of the model require further validation through repeated cross-validation in future research.

Global SHAP Interpretation

Mean Absolute SHAP Value analysis was applied to evaluate the overall variable impact on the XGBoost decision architecture.

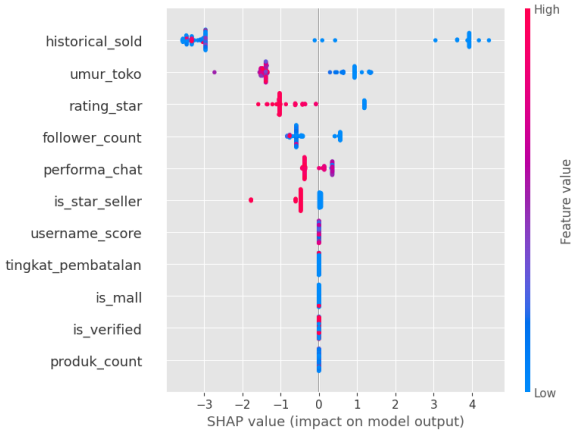


Figure 6. SHAP summary plot for global feature interpretation.

This visualization produced three dominant features: historical_sold at 48.42%, umur_toko at 18.73%, and rating_star at 14.76%. To

The work is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i4.651
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

obtain these percentages, the absolute SHAP values for each feature were aggregated and normalized against the total sum of all absolute SHAP values across the dataset, allowing for an intuitive relative comparison.

Table 4. Three features with the largest SHAP contributions.

Ranking	Features	SHAP Contribution	Observed Significance
1	historical_sold	48.42%	Sales history is the most dominant factor; high values tend to support a "Safe" classification.
2	umur_toko	18.73%	Longer operational duration tends to lower the probability of fraud.
3	rating_star	14.76%	Reputation ratings help distinguish between legitimate and fake stores.

These results have practical implications: the model primarily reads signals representing operational maturity and store reputation. This finding aligns with the seller-centric orientation of the research, but must not be interpreted as a causal relationship. SHAP values explain model behavior towards features, not proving that a feature singly causes a store to become Fraud.

Local SHAP Interpretation

On testing instance with index 35, the XGBoost model produced a SAFE prediction with confidence reaching 99.89%.

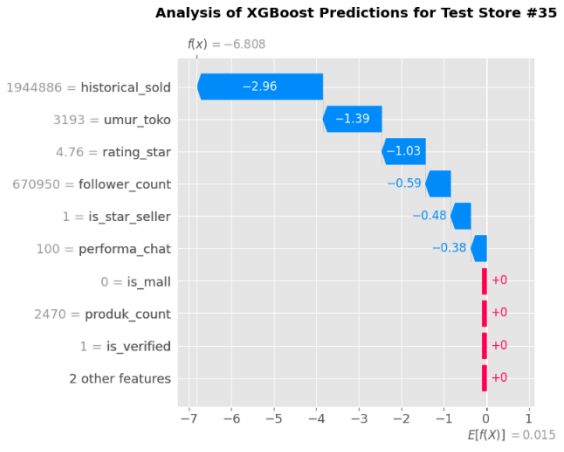


Figure 7. SHAP waterfall plot for the store instance at index 35.

The visualization demonstrates how SHAP translates complex model decisions into readable explanations by quantifying both the magnitude and direction of each feature's effect. For this specific instance, the negative SHAP values push the model's output strongly toward the Safe (Class 0) classification. The three most significant factors driving this specific justification are historical_sold (1,944,886 sold products), umur_toko (8 years, 9 months, 10 days), and rating_star (4.76). While this single Safe instance serves as a primary representative example to illustrate the prototype's real-time translation capability, the same underlying SHAP mechanics apply inversely to Fraud instances, where positive SHAP values would drive the risk indication. Local interpretation is crucial because end-users require transparent, instance-specific rationales alongside the predicted risk labels.

Prototype Implementation

The XGBoost model, SHAP explainer, and feature column sequence were serialized using the joblib library. The Flask backend, hosted on Hugging Face Spaces, receives the store profile payload, executes the identical feature transformations used during the training phase, generates predictions with confidence scores, and extracts the top three driving features via SHAP. On the client side, these components are seamlessly integrated into a Manifest V3 standard Chrome Extension.

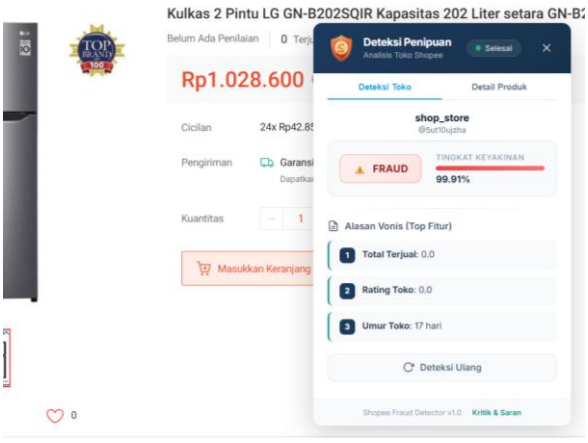


Figure 8. Example of a Chrome Extension prototype interface for store profile detection.

The deployment of this prototype demonstrates the successful downstreaming of the predictive model into a functional fraud-risk screening tool. However, acknowledging its

The work is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License

reliance on heuristic labeling and the absence of a controlled user study to empirically measure its impact on reducing consumer losses, the system's current utility is strictly bounded. Furthermore, because this architecture operates externally without integration into Shopee's official internal database, the prototype must be explicitly positioned as a research demonstrator for client-side early-warning risk indication, rather than an absolute legal determination of fraud.

(the prototype is accessible at: https://chromewebstore.google.com/detail/mjomfkbmbkfhdchhndefacelbmacdhco?utm_source=item-share-cb)

CONCLUSIONS AND SUGGESTIONS

Conclusion

This study successfully produced a machine learning classification model for indicating fraudulent sellers on the Shopee platform based on store profile metadata, utilizing the XGBoost algorithm integrated with SHAP. From an initial extraction of 460 profiles, 457 clean observations were validated (355 Safe and 102 Fraud) comprising 11 predictor features. Utilizing an 80:20 stratified data split and a scale_pos_weight parameter of 3.48 to handle class imbalance, the evaluation on 92 test instances yielded an Accuracy of 95.65%, Precision of 84.00%, Recall of 100.00%, and an F1-Score of 91.30%. The 100.00% Recall achievement demonstrates that all Fraud instances within the specific test set were successfully detected, while the 84.00% Precision indicates the presence of minor false positives. Consequently, this system's output must be strictly positioned as an early-warning risk indication rather than an absolute legal determination of fraud status.

SHAP analysis revealed that historical_sold, umur_toko, and rating_star are the three most dominant predictive features, contributing 48.42%, 18.73%, and 14.76%, respectively. Local interpretation, exemplified by store index 35, demonstrated how specific quantitative values (historical_sold of 1,944,886 items, a store age of 8 years 9 months 10 days, and a rating_star of 4.76) synergistically drove the model to output a SAFE prediction with 99.89% confidence. This classification pipeline was successfully downstreamed into a client-side Chrome Extension prototype connected to a Flask backend. Ultimately, the main contribution of this research lies in establishing a proactive, transparent, and seller-centric fraud detection framework utilizing publicly accessible metadata.

Suggestion

While this study successfully established a proof-of-concept risk screening tool, several methodological limitations must be acknowledged. These include a relatively small purposive sample size, a manual heuristic labeling process reliant on community consensus, and the absence of extensive software usability testing. To address these boundaries, future research is advised to:

1. Expand the sample size and observation period, employing repeated cross-validation to ensure broader model generalization and stability across diverse data splits.
2. Strengthen label validity by incorporating formal inter-rater reliability metrics (e.g., Cohen's Kappa) involving specialized e-commerce security domain experts.
3. Conduct comprehensive software evaluations on the prototype, including API latency, cross-browser compatibility, and resilience against continuous changes in platform webpage structures.

REFERENCES

Adams, V. (2024). Graph Neural Networks Analytics and AI for Fraud Detection. https://doi.org/10.64206/vvc03q65

Ardhitha, R., Anugerah, R., & Sutabri, T. (2025). Analisis Penerapan Machine Learning dan Algoritma Anomali untuk Deteksi Penipuan pada Transaksi Digital. *Repeater: Publikasi Teknik Informatika dan Jaringan, 3*(1), 80–90. https://doi.org/10.62951/repeater.v3i1.345

Awalia, S. A., Jedha Arhansyah, R., Rahadian Adi Nugroho, F., & Amzar Thapsuandji, A. (2024). Perlindungan Hukum terhadap Konsumen dalam Transaksi E-commerce Shopee. Jurnal Potensial, 3(1). https://jurnalilmiah.org/journal/index.php/potensial/article/view/740

Awosika, T., Shukla, R. M., & Pranggono, B. (2023). Transparency and Privacy: The Role of Explainable AI and Federated Learning in Financial Fraud Detection. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3394528

Bucci, V., & Li, W. (2023). From Manifest V2 to V3: A Study on the Discoverability of Chrome Extensions. *Springer*. https://doi.org/10.1007/978-3-031-49187-0_10

The work is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License

Chen, F. (2024). Research on Real-time E-commerce Price Comparison System Using Python Web Scraping Technology. *International Journal of Computer Science and Information Technology, 4*(2), 127–136. https://doi.org/10.62051/ijcsit.v4n2.18

Dekou, R., Savo, S., Kufeld, S., Francesca, D., & Kawase, R. (2022). Machine Learning Methods for Detecting Fraud in Online Marketplaces. *arXiv preprint*. https://doi.org/10.48550/arXiv.2205.00000

Farsi, S., & Chowdhury, M. (2025). EcomFraudEX: An Explainable Machine Learning Framework for Victim-Centric and Dual-Sided Fraud Incident Classification in E-Commerce. *ICST Transactions on Scalable Information Systems, 12*. https://doi.org/10.4108/eetsis.6789

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *arXiv preprint*. https://doi.org/10.48550/arXiv.2207.08815

Hertati, L., & Yadiati, W. (2024). Exploring Human Capital dalam Tingkat Pendidikan Peran Aplikasi Sistem Informasi Akuntansi Mengatasi Deteksi Fraud pada Aplikasi Shopee. *Jurnal Riset Akuntansi, 16*(1). https://ojs.unikom.ac.id/index.php/jira/article/download/11457/4283/43018

Karim, A. A. (2025). Analisis Prediktif untuk Mendeteksi Penipuan E-Commerce Menggunakan Algoritma Pembelajaran Mesin. The Indonesian Journal of Computer Science, 14(2). https://doi.org/10.33022/ijcs.v14i2.4202

Kaya, O., Shahen Shah, A. F. M., Karabulut, M. A., Karahan, S. N., Osmanca, M. S., & Acir, N. (2026). Explainable Artificial Intelligence (XAI): Concepts, Applications, Challenges, and Future Perspectives. *IEEE Access*. https://doi.org/10.1109/ACCESS.2026.3663161

Kustanto, P. (2024). Implementation of a Machine Learning with a Python Programming Approach to Predict the Level of Disease Cases. *Global Conference on Innovation and Research (GCIR), 1*(1). https://doi.org/10.38035/gcir.v1i1

Lhaura, L., Khairul Anam, M., Firdaus, M. B., Yunefri, Y., & Rahmi, N. A. (2024). Enhancing Machine Learning Model Performance in Addressing Class Imbalance. *COGITO Smart Journal, 10*(1). https://cogito.unklab.ac.id/index.php/cogito/article/download/626/327

Maresti, F. A., Anugraheni, G. M., Hargiyanto, R. A., & Mustaqim, K. (2024). Penerapan Exploratory Data Analysis (EDA) dan Analisis Recency, Frequency, and Monetary (RFM) untuk Segmentasi Pelanggan E-Commerce. *Jurnal Competitive*. https://ejurnal.ulbi.ac.id/index.php/competitive/article/download/4059/1405

Miruna, S., Amali, J., Dharshini, K., & Iswarya, A. (2026). Fraud Detection in Online Shopping Using XGBoost. *International Journal for Multidisciplinary Research (IJFMR), 7*(5). https://www.ijfmr.com/papers/2025/5/58724.pdf

Mohammed, Z. H., Hatem Khorsheed, F., & Ahmed, G. J. (2025). Ensemble Deep Learning Strategy for Handling Imbalanced Credit Card Fraud Data. *Journal of Informatics and Computer Science (JOINCS), 8*(2). https://joincs.umsida.ac.id/index.php/joincs/article/download/1670/1858

Ponmozhi, K., & Surya, J. (2026). ExhibitorIQ: Intelligent Exhibitor Data Extraction and Event Outreach Platform. *Iconic Research and Engineering Journals, 9*(9). https://doi.org/10.64388/IREV9I9-1715669

Rahim, N., & Basheer, M. K. P. (2025). A hybrid explainability framework for recommender systems using SHAP with natural language interpretations. *International Journal of Innovative Research and Scientific Studies, 8*(6), 687–703. https://doi.org/10.53894/ijirss.v8i6.9669

Ranglani, H. (2024). Empirical Analysis of the Bias-Variance Tradeoff Across Machine Learning Models. *Machine Learning and Applications: An International Journal, 11*(4), 1–12. https://doi.org/10.5121/mlaij.2024.11401

Raufi, B., Finnegan, C., & Longo, L. (2024). A comparative analysis of SHAP, LIME, ANCHORS, and DICE for interpreting a dense neural network in credit card fraud detection. *Communications in Computer and Information Science, 2156*. https://doi.org/10.1007/978-3-031-63803-9_20

Salsabila, L., & Hutagalung, F. S. (2025). Performance Comparison of Random Forest and XGBOOST Algorithms on E-Commerce App Sentiment Analysis. *Indonesian Journal of Applied Technology, 2*(2), 57–67. https://jurnal.umsu.ac.id/index.php/Ijatcos/article/download/26211/pdf

Shimaoka, A. M., Ferreira, R. C., & Goldman, A. (2024). The evolution of CRISP-DM for Data Science: Methods, Processes and

The work is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License

Frameworks. *SBC Reviews on Computer Science, 4*(1), 28–43. https://doi.org/10.5753/reviews.2024.3757

Stafeev, A., & Pellegrino, G. (2024). SoK: State of the Krawlers - Evaluating the Effectiveness of Crawling Algorithms for Web Security Measurements. *33rd USENIX Security Symposium (USENIX Security 24)*. https://www.usenix.org/conference/usenixsecurity24/presentation/stafeev

Tanasă, A.-M., Oprea, S.-V., & Bâra, A. (2023). Web Scraping and Review Analytics. Extracting Insights from Commercial Data. *Ovidius University Annals, Economic Sciences Series, 23*(2). https://doi.org/10.61801/OUAESS.2023.2.45

Teoh, T. T., & Rong, Z. (2022). *Python for Artificial Intelligence*. Springer. https://doi.org/10.1007/978-981-16-8615-3_1

Thanathamathee, P., Sawangarreerak, S., Chantamunee, S., & Nizam, D. N. M. (2024). SHAP-Instance Weighted and Anchor Explainable AI: Enhancing XGBoost for Financial Fraud Detection. *Emerging Science Journal, 8*(6), 2404–2430. https://doi.org/10.28991/ESJ-2024-08-06-016

Wijaya, W. (2025). Exploratory Data Analysis Pada Dataset Ulasan Pelanggan Mc'Donald di Australia. *Journal of Smart Technology and Engineering, 1*(1). https://doi.org/10.28932/jste.v1i1.13176

Wu, C.-Y., Kuo, C.-C., & Yang, C.-S. (2023). Phishing Detection with Browser Extension Based on Machine Learning. *18th Asia Joint Conference on Information Security (AsiaJCIS)*, 81–87. https://doi.org/10.1109/AsiaJCIS60284.2023.00023

Yin, A., & Li, B. (2025). XGBoost-Based Fraud Detection and Risk Assessment in E-commerce Transactions. *Proceedings of 2025 2nd International Conference on Big Data and Digital Management (ICBDDM 2025)*, 505–511. https://doi.org/10.1145/3768801.3768881

Zheng, A., & Casari, A. (2024). *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media.

The work is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License