Predicting Fraud Cases in E-Commerce Transactions Using Random Forest Regression: A Data Mining Approach for Enhancing Cybersecurity and Transaction Integrity

Yusuf Durachman[1,*], Abdul Wahab Bin Abdul Rahman[2,]

[1]State Islamic University Syarif Hidayatullah, Jakarta, Indonesia

[2]International Islamic University Malaysia, Kuala Lumpur, Malaysia

ABSTRACT

Fraudulent activities in e-commerce pose significant risks to businesses and consumers alike, resulting in financial losses and eroding trust in online transactions. This study aims to address this issue by developing a predictive model for fraud cases using Random Forest Regression, a robust machine learning technique known for handling nonlinear relationships and high-dimensional data. The dataset comprises daily transaction metrics such as fraud cases, transaction errors per million, transparency rating, security incidents, cyber attacks, audit compliance scores, transaction speeds, and customer trust indices, collected over multiple years. The methodology involves extensive data preprocessing, including temporal feature extraction from date information, and exploratory data analysis to identify key relationships among features. Correlation analysis revealed that transaction errors per million and security incidents are highly correlated with fraud cases, serving as important predictors. The dataset was split into training and testing sets, with the Random Forest model trained on 80% of the data and evaluated on the remaining 20%. Results indicate that the Random Forest model predicts fraud cases with high accuracy, achieving an R-squared score of 0.9832 and low error metrics (MAE of 21.07 and RMSE of 26.26). Feature importance analysis identified transaction errors per million as the most influential variable, confirming its critical role in fraud detection. Despite these promising results, limitations such as potential data imbalance and model interpretability challenges remain and warrant further research. This research contributes to the growing body of knowledge applying machine learning to cybersecurity and fraud detection, demonstrating practical applicability for improving e-commerce transaction security. The findings also have implications for cyberlaw, suggesting that advanced predictive tools can enhance regulatory enforcement and help develop more secure online commerce environments. Future work will explore incorporating additional features and alternative algorithms to further improve model robustness and transparency.

Keywords Fraud Detection, E-Commerce Security, Random Forest Regression, Machine Learning, Cyberlaw

Submitted 4 February 2025
Accepted 18 April 2025
Published 3 June 2025

*Corresponding author
Yusuf Durachman,
yusuf.rahman@uinjkt.ac.id

Additional Information and Declarations can be found on page 128

DOI: 10.63913/jcl.v1i2.6

Introduction

The prevalence of fraud in e-commerce transactions has emerged as a significant concern, posing threats to both online businesses and their users. The evolution of technology and the increasing reliance on digital platforms for transactions have facilitated various forms of fraud, making it imperative to develop robust detection mechanisms. Notably, the challenges inherent in

How to cite this article: Y.Durachman and A.W.B.A.Rahman, “Predicting Fraud Cases in E-Commerce Transactions Using Random Forest Regression: A Data Mining Approach for Enhancing Cybersecurity and Transaction Integrity,” *J. Cyber. Law.*, vol. 1, no. 2, pp. 116-130, 2025.

detecting e-commerce fraud are manifold, primarily revolving around data quality and the scarcity of reported fraud incidents.

A fundamental challenge in detecting e-commerce fraud is the imbalance in data distribution between valid and fraudulent transactions. The frequency of valid transactions generally overshadows that of fraudulent ones, resulting in poor data quality that hampers the effectiveness of machine learning algorithms, as discussed by [1] and further supported by [2]. This issue is exacerbated by the prevalence of unreported incidents, where victims often lack awareness of where to report fraud or harbor doubts about the efficacy of legal remedies, as highlighted by [3], [4]. The clandestine nature of such activities makes it further difficult for analytic models to be trained on the existing data. Research [5] emphasizes the need for comprehensive detection systems to address the challenges posed by sparse data.

Recent advancements in machine learning and artificial intelligence present new opportunities for improving fraud detection systems across e-commerce platforms. For instance, sophisticated techniques, such as sentiment analysis, have been proposed to identify potentially fraudulent behaviors by analyzing user sentiments associated with transactions [6]. Furthermore, methodologies involving anomaly detection are proving increasingly popular, as they enhance capabilities to flag irregular transactional activities that deviate from established patterns [7], [8]. The integration of these technologies into existing frameworks could significantly bolster the defense against e-commerce fraud, allowing for a more proactive approach in identifying potential threats before they escalate.

Fraud detection in e-commerce transactions presents numerous challenges primarily due to the dynamic and evolving nature of fraudulent activities. As methods employed by fraudsters become increasingly sophisticated, traditional detection mechanisms often fall short. This evolving landscape necessitates the adaptation of innovative technological frameworks to enhance fraud detection in e-commerce settings. One key challenge is related to the class imbalance often observed between legitimate and fraudulent transactions. As Prasad [9] articulates, the number of fraudulent transactions (a minority class) is generally much lower than the number of valid ones (a majority class). This imbalance can significantly undermine the effectiveness of machine learning models trained on historical transaction data, as they may not adequately learn from the scarce examples of fraud. To address this issue, methods like enhanced oversampling techniques have been proposed to improve detection efficiency by better representing the minority class in training datasets [9].

Machine learning and deep learning techniques offer promising avenues for addressing these challenges. For instance, Damayanti and Adrianto [1] report on the success of Hidden Markov Models (HMM) in identifying fraudulent activity, achieving high accuracy across different transaction types using machine learning. Similarly, Rout and Jaiswal [10] highlight the robustness of deep learning methodologies in extracting complex patterns from data, enabling them to adapt to and detect evolving fraud strategies in real-time. The ability of these systems to continuously learn and update is critical in a landscape where fraud tactics are rapidly changing. Furthermore, Hasugian and Suharjito [11], [12] elucidates the applicability of outlier detection approaches, identifying fraudulent transactions as anomalies based on atypical behaviors, such as unusually high transaction values or frequencies that deviate from a customer's usual behavior pattern. This aligns with the necessity for sophisticated anomaly detection strategies, as discussed by Du et al [7], who categorize methods into rule-based, machine learning-based, and graph-based strategies. Rule-based systems, while foundational, struggle against adaptive measures taken by fraudsters, highlighting the need for an integrated approach that leverages multiple detection algorithms.

The main objective of this study is to develop a predictive model for fraud cases in e-commerce transactions using the Random Forest Regression algorithm. By leveraging transaction data and related features, the study aims to accurately forecast the number of fraud occurrences. This prediction capability is intended to support enhanced cybersecurity measures by enabling timely detection and prevention of fraudulent activities. This research is significant because fraud remains one of the biggest challenges facing e-commerce platforms, often resulting in financial losses and diminished customer trust. By improving fraud detection systems through machine learning, the study offers a practical approach to safeguard online transactions. Enhancing transaction integrity is vital for maintaining the confidence of both consumers and merchants in digital marketplaces. Furthermore, the findings of this study contribute to the ongoing development of data-driven security solutions. The use of Random Forest Regression allows for robust modeling of complex patterns in transactional data, which can help organizations proactively manage fraud risks. This contributes not only to cybersecurity efforts but also to the broader domain of e-commerce regulation and compliance.

Literature Review

E-Commerce Fraud and Cybersecurity Challenges

E-commerce fraud is an increasingly complex phenomenon characterized by various fraudulent activities such as phishing, account takeover, and transaction manipulation. These methods not only threaten the integrity of online transactions but also significantly undermine consumer trust and business viability. Understanding these challenges requires a multidimensional approach that combines technological innovation with consumer education and robust cybersecurity measures. Phishing is one of the most prevalent forms of e-commerce fraud, employing deceptive techniques to trick users into providing sensitive information. Phishers typically create counterfeit websites that closely mimic legitimate ones, utilizing similar designs and layouts to elicit financial and personal data from unsuspecting users [13]. Taha [13] notes that the rapid increase in online user engagement due to convenience and flexibility has made e-commerce platforms attractive targets for phishing attacks, complicating the security landscape further. The necessity of a robust phishing detection system is underscored by the increasing sophistication of these attacks, which often use tailored pages designed to look legitimate to deceive users [14].

Account takeovers present another prominent challenge in e-commerce fraud. This form of fraud occurs when malicious actors gain unauthorized access to user accounts, often through methods like credential stuffing or social engineering. The implications are severe, leading to unauthorized transactions, altered account settings, and breaches of personal information. As Rout and Jaiswal [10] observe, deep learning models are increasingly employed to analyze various indicators associated with abnormal transaction patterns and user behaviors, effectively detecting such fraudulent activities. This technology allows businesses to identify potential account takeover attempts in real-time by scrutinizing transaction histories and login patterns, thus acting before significant damage occurs. Moreover, the growing threat landscape necessitates not only technological advancements but also comprehensive regulatory frameworks to protect e-consumers against fraud. Razali et al [4] emphasize the importance of legal protections for e-consumers to mitigate the risks associated with various forms of e-commerce fraud. These protections can discourage fraudulent activities and bolster consumer confidence in online transactions, thereby fostering a more secure e-commerce environment.

Data Mining in Cybersecurity

Data mining techniques have become vital tools in the detection of fraudulent transactions in e-commerce and finance. Among these techniques, classification and regression algorithms stand out due to their ability to analyze large datasets, identify patterns, and uncover anomalies indicative of fraud. The increasing sophistication of fraudulent schemes necessitates the adoption of advanced data mining methodologies to enhance the accuracy and effectiveness of fraud detection systems. Classification algorithms are particularly useful in categorizing transactions as either legitimate or fraudulent. Gupta [15] highlights that various classification techniques, including decision trees, support vector machines (SVM), and neural networks, have been successfully employed in fraud detection scenarios. These techniques utilize historical data to learn the distinguishing features of legitimate transactions compared to fraudulent ones, enabling systems to classify new transactions accurately. For instance, logistic regression has proven effective in various financial contexts, and its adaptability makes it one of the preferred choices for initial fraud detection models [16].

Regression algorithms further complement classification approaches by predicting the likelihood of fraud based on transactional attributes. For example, Dastjerdi et al [17] emphasize the capability of regression modeling to identify patterns in managerial reports that signal potential high fraud risk. By integrating text mining techniques with regression analysis, businesses can detect subtle signs of fraud based on unstructured data, further enhancing their detection frameworks. The role of clustering algorithms is crucial, as these techniques help in segmenting transactions into groups based on similarities, which can aid in identifying outliers that may signify fraudulent activity. Cho [18] indicates that clustering techniques can enhance the detection of anomalies when applied to large datasets, revealing hidden correlations that traditional single-method algorithms might miss. This capacity underscores the value of using a hybrid approach, combining different data mining techniques to capture a more comprehensive view of potential threats.

Fraud Detection Models and Algorithms

The effectiveness of various machine learning algorithms, particularly Random Forest, SVM, and Decision Trees, has been validated in the field of fraud detection within financial transactions. Each of these algorithms exhibits unique strengths, enabling robust detection systems capable of addressing the complexities of evolving fraudulent activities. Random Forest is a widely utilized ensemble method that enhances classification accuracy through the aggregation of multiple decision trees. Meghana and R [19] demonstrate the efficacy of a novel Random Forest algorithm in detecting fraudulent service enrollment websites, showcasing its superior accuracy compared to other machine learning algorithms like XGBoost. Furthermore, a study specifically comparing Random Forest to logistic regression in mobile money transactions reports that Random Forest outperforms logistic regression, achieving accuracy rates of 98%. This highlights Random Forest's capability to process complex datasets and identify patterns indicative of potential fraud.

SVM also play a crucial role in fraud detection, yielding impressive results due to their effectiveness in handling high-dimensional spaces. Research by Cho [18] indicates that SVMs have been successfully applied in supervised learning contexts, demonstrating superior performance over traditional classifiers in identifying fraudulent cases. The effectiveness of SVMs stems from their ability to create hyperplanes that effectively separate legitimate and fraudulent transactions, thus enhancing the accuracy of fraud detection models. Decision Trees are another foundational component in fraud detection algorithms, known for their transparency and interpretability. These models work by splitting data into branches based on feature values, leading to a decision outcome that categorizes transactions as either fraudulent or legitimate. Mousa [16] illustrates how Decision Trees, in conjunction with other classifiers like logistic regression and neural networks, have been employed to uncover hidden relationships in financial datasets, significantly contributing to the identification of fraudulent transactions. The simplicity of Decision Trees aids in quick model interpretation, making them particularly valuable in operational settings where swift decisions are crucial. The integration of these algorithms into hybrid models further enhances detection systems. For instance, the use of ensemble methods that combine predictions from multiple models leads to greater accuracy and reliability. Pk's [20] introduces a Bayesian optimized Random Forest classifier that integrates advanced feature analysis and real-time data adaptation, reflecting the increasing trend of combining multiple machine learning techniques for improved detection of credit card fraud.

Random Forest Regression for Fraud Prediction

Random Forest Regression is a robust machine learning technique that has emerged as an effective tool for detecting fraud in financial transactions, particularly due to its capacity to manage imbalanced datasets and identify complex fraud patterns. This algorithm's strength lies in its ensemble approach, which utilizes multiple decision trees to enhance predictive accuracy and reduce overfitting, making it suitable for the nuanced task of fraud detection. One of the key advantages of the Random Forest algorithm is its effectiveness in handling imbalanced datasets, a common scenario in fraud detection where fraudulent transactions are significantly outnumbered by legitimate ones. Lokanan's study [21] highlights the consistency of the Random Forest model compared to other classifiers in predicting mobile money transaction fraud, achieving high precision and recall rates. This capability is crucial as it allows the model to learn from a diverse set of features associated with both legitimate and fraudulent transactions while mitigating the risk of bias towards the majority class.

The ability of Random Forest Regression to analyze the importance of various features is also noteworthy. Lucas et al [22] discuss the potential future work of combining predictions from various models, including Random Forest, to enhance fraud detection accuracy, although their study does not directly validate this approach's effectiveness in practice. Moreover, research by Xu et al [23] illustrates the adaptability of Random Forest. They found that combining Random Forest with other algorithms like Support Vector Machines enhances fraud prediction models, highlighting how different methods can outperform others in certain contexts. This adaptability makes Random Forest suitable for varied fraud detection scenarios, from e-commerce to credit card fraud. Furthermore, the model's capability for real-time analysis enhances operational responsiveness to potential fraudulent activities. By continuously learning from new transaction data, Random Forest Regression not only identifies existing fraud patterns but also adapts to emerging trends, which is crucial in an environment where fraudulent tactics evolve rapidly.

Method

Figure 1 illustrates our research methodology, which follows a sequential five-step process beginning with data collection, moving through preparation and analysis, model training, and evaluation, and concluding with the saving of the final results.

Data Collection
Data Preparation & Analysis
Model Training
Model Evaluation
Results & Saving

Figure 1 Research Method Flowchart

Data Collection and Loading

The dataset for this study was obtained from an Excel file titled "fraud and blockchain.xlsx", containing multiple transaction-related features relevant to e-commerce fraud detection. These features include daily counts of fraud cases, transaction errors per million, transparency ratings, security incidents, cyber attacks reported, audit compliance scores, transaction speeds, and customer trust indices. The data was imported using the pandas library in Python. Initial inspection involved checking the dataset's shape, data types, and completeness to ensure its suitability for subsequent analysis and modeling.

Data Preprocessing and Feature Engineering

Preprocessing began with converting the date column from a string format into a datetime object to enable temporal feature extraction. From the date column, additional features such as year, month, day, dayofweek (representing the day of the week with Monday=0), and dayofyear were derived, expanding the dataset's temporal dimension. This step helps the model to capture potential seasonal or day-based patterns in fraud occurrence. The original date column was then set as the dataframe index to facilitate time-series visualization if needed. Only numerical features were selected for modeling to ensure compatibility with machine learning algorithms, and any missing values or inconsistencies were checked and addressed.

Exploratory Data Analysis (EDA) and Visualization

EDA involved examining the distribution of each key feature using histograms combined with kernel density estimates to detect skewness, outliers, or multimodal patterns. Features like Fraud Cases, Transaction Errors per Million, and Security Incidents were individually visualized to understand their behavior over the dataset's timeline. A Pearson correlation matrix was computed to assess linear relationships between all numerical variables, with a heatmap visualization highlighting positive and negative correlations. Special focus was given to the correlations with the target variable Fraud Cases to identify the most influential predictors. Additionally, scatter plots were generated for features with correlation coefficients above 0.1 in absolute value to visually confirm relationships and potential predictive power.

Random Forest Regression

The dataset was split into training and testing subsets using an 80:20 ratio, with the train_test_split function and a fixed random_state of 42 to ensure consistent splits across runs. The Random Forest Regressor was selected for its robustness against overfitting, ability to handle nonlinear relationships, and suitability for datasets with mixed feature types. The model was initialized with 100 decision trees (n_estimators=100), providing a balance between prediction accuracy and computational efficiency. The maximum depth of each tree was limited to 10 (max_depth=10) to prevent overfitting by restricting the complexity of individual trees. The minimum number of samples required to split an internal node was set to 5 (min_samples_split=5), ensuring that splits occur only when sufficient data points exist, which helps maintain model generalization. Parallel computation was enabled via n_jobs=-1 to use all available CPU cores for faster training.

Model Evaluation and Interpretation

After training, model performance was evaluated on the test set using multiple metrics: Mean Absolute Error (MAE), which measures average absolute prediction errors; Mean Squared Error (MSE), which penalizes larger errors more heavily; Root Mean Squared Error (RMSE), providing error magnitude in the original units; and R-squared (R^2), indicating the proportion of variance explained by the model. These complementary metrics provide a thorough assessment of predictive accuracy and model fit. To visually assess prediction quality, a scatter plot of actual vs predicted fraud cases was created, with a diagonal reference line representing perfect predictions. Feature importance scores extracted from the trained Random Forest model were analyzed to rank features by their contribution to predictions, aiding interpretability and identifying key drivers of fraud risk.

Model Saving and Output Management

For reproducibility and practical use, the trained Random Forest model was serialized and saved using the joblib library, allowing the model to be efficiently loaded for future inference without retraining. All generated plots—distribution histograms, correlation heatmaps, scatter plots, actual vs predicted comparison, and feature importance bar charts—were saved as image files in organized folders. Evaluation metrics were also exported as CSV files for documentation and reporting. This comprehensive saving strategy ensures easy access for further analysis, presentation, or deployment in fraud detection workflows.

Result and Discussion

Dataset Overview and Initial Inspection

The dataset used for this study consists of 1,036 daily records spanning multiple years, with 9 primary features including the target variable, Fraud Cases. Initial data inspection confirmed there were no missing values, and all columns had appropriate data types—ranging from integer and float to datetime for the date column. The first five rows displayed daily fraud cases ranging from 911 to 970 and varying transaction errors and security-related metrics. After preprocessing, additional date-related features such as year, month, day, day of week, and day of year were extracted to enrich the dataset for temporal pattern recognition.

Descriptive Statistics and Exploratory Data Analysis (EDA)

Descriptive statistics showed that the average number of fraud cases was approximately 697, with a standard deviation of 208, indicating considerable day-to-day variation. Transaction errors per million averaged around 331, and security incidents averaged 310 per day. Notably, temporal features such as year and month captured data from 2020 to 2022, allowing temporal trend analysis. EDA visualizations revealed the distribution of each feature, highlighting some skewness in fraud cases and transaction errors. A correlation heatmap demonstrated strong positive correlations between fraud cases and transaction errors per million (0.97), as well as with security incidents (0.92). Interestingly, transparency rating and year showed strong negative correlations with fraud cases (-0.91 and -0.96 respectively), suggesting improvements over time or possibly data collection artifacts. Scatter plots between fraud cases and top correlated features visually confirmed these relationships.

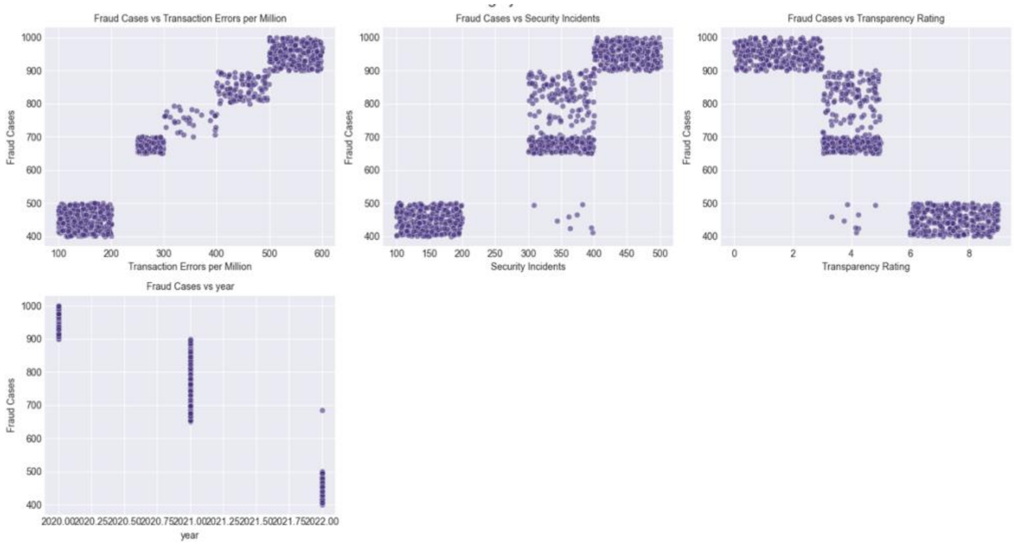

Figure 2 **Scatter Plots of Key Features**

Figure 2 illustrates the relationship between fraud cases and several key features. The plot comparing fraud cases with transaction errors per million reveals a strong positive correlation, where higher transaction error rates align with increased fraud cases. This indicates that transaction errors are a reliable indicator of potential fraud activity. Similarly, the scatter plot of fraud cases versus security incidents also shows a positive correlation, although with more variability, suggesting that days with more security incidents tend to have more fraud cases, but other factors might influence this relationship. Interestingly, the plot of fraud cases against transparency rating displays a negative correlation, where higher transparency ratings correspond to fewer fraud cases, implying that increased transparency could be linked to better fraud prevention. Lastly, the scatter plot of fraud cases versus year indicates a downward trend over time, with fraud cases generally decreasing from 2020 to 2022, possibly reflecting improvements in security measures or data collection.

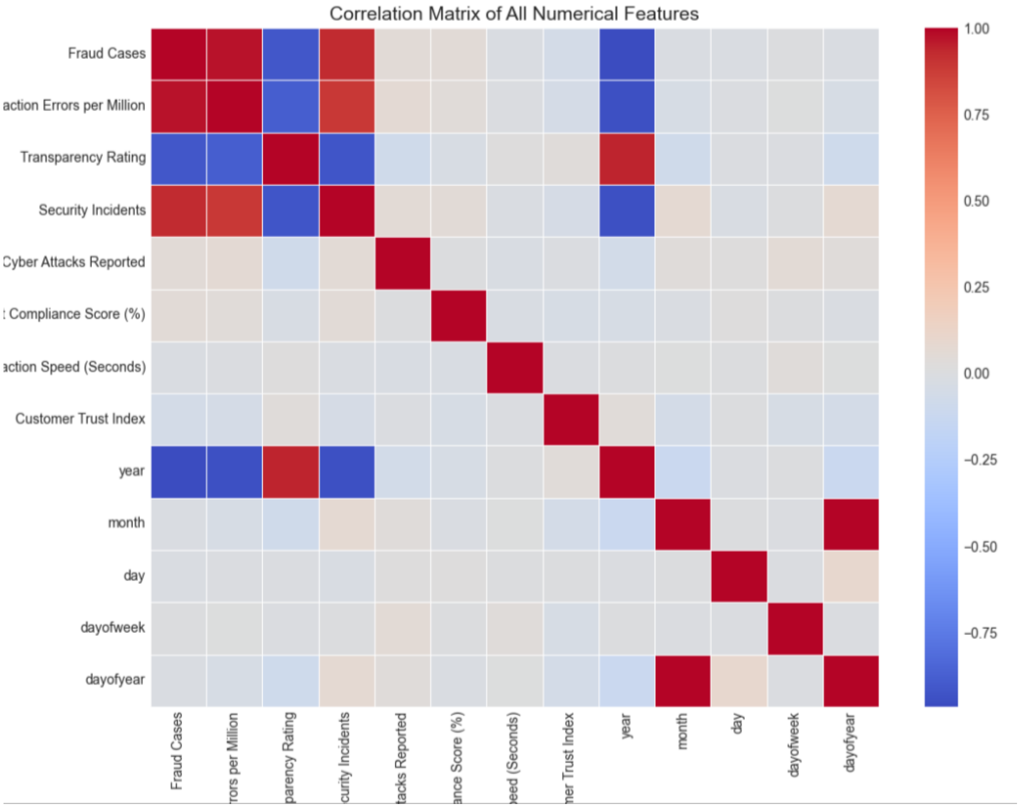

Figure 3 Correlation Matrix

Figure 3 visually summarizes the linear relationships between all numerical variables. Darker reds indicate strong positive correlations, while blues indicate strong negative correlations. Fraud cases show very high positive correlations with transaction errors per million and security incidents, reinforcing their importance as fraud predictors. Conversely, transparency rating and year exhibit strong negative correlations with fraud cases, suggesting that as transparency and time increase, fraud cases tend to decline. Other features, such as cyber attacks reported and audit compliance score, display weak or no clear correlations with fraud cases, indicating they may have less predictive power in this dataset. The heatmap also highlights strong internal correlations among date-derived features such as year, month, day, and day of the year, reflecting their inherent temporal relationships.

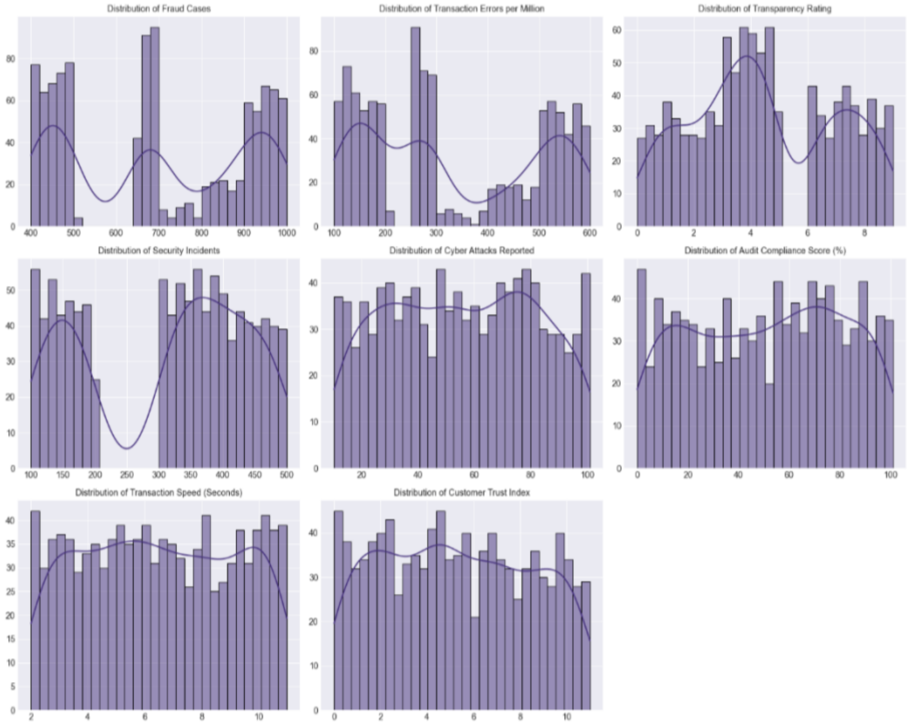

Figure 4 Distribution Plots of Key Features

Figure 4 provide insight into the frequency and spread of values for each key feature. Fraud cases show a bimodal distribution with peaks around 450 and 900 cases, indicating clusters of low and high fraud activity days. Transaction errors per million exhibit a similar bimodal pattern, aligning with the strong correlation to fraud cases. Transparency rating and audit compliance score show relatively even distributions across their ranges but with slight multimodal tendencies, suggesting periods of varying transparency and compliance. Security incidents and cyber attacks reported present somewhat uniform distributions, implying consistent daily occurrence rates. Transaction speed and customer trust index distributions appear relatively normal with mild variations, indicating stable transaction processing times and trust levels across the dataset.

Model Training and Performance

The data was split into 80% training and 20% testing sets, yielding 828 training samples and 208 test samples. A Random Forest Regressor was trained with 100 trees, a maximum depth of 10, and a minimum split of 5 samples per node. This configuration balanced model complexity with generalization capacity. On the test set, the model achieved excellent predictive performance, with an R-squared value of 0.9832, indicating that over 98% of the variance in fraud cases was explained by the model. The MAE was 21.07, and the RMSE was 26.26, reflecting low average prediction errors relative to the fraud case scale. The scatter plot of actual versus predicted fraud cases showed points closely aligned to the ideal diagonal line, confirming high model accuracy.

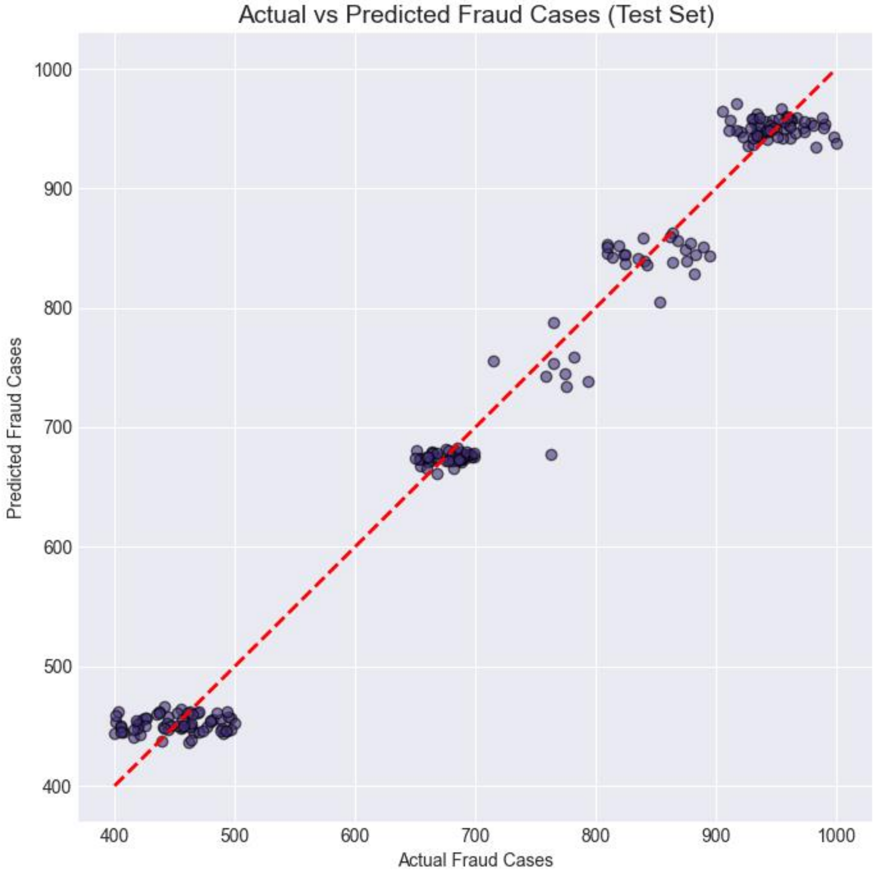

Figure 5 Actual vs Predicted Values

Figure 5 illustrates the comparison between actual and predicted fraud cases on the test set, showcasing the model's predictive accuracy. Each point on the scatter plot represents a daily record, with the x-axis displaying the actual observed fraud cases and the y-axis showing the model's predicted values. The red dashed diagonal line represents the ideal scenario where predicted values perfectly match actual values. Most data points cluster tightly around this diagonal, indicating that the Random Forest model is highly accurate in forecasting fraud cases. This close alignment reflects the model's strong ability to generalize from training data to unseen data, with minimal prediction error even across varying fraud levels.

Feature Importance Analysis

Analysis of feature importance revealed that Transaction Errors per Million was by far the most influential predictor, accounting for over 93% of the model's decision-making process. This confirms the intuitive and statistical relationship between transaction errors and fraud occurrence. The year feature contributed approximately 3.4%, reflecting temporal trends, while other features such as Security Incidents, Transparency Rating, and Cyber Attacks Reported had relatively minor contributions, each less than 1%. This distribution of importance suggests that, for this dataset, transactional error frequency is the key indicator for predicting fraud cases, and temporal factors also provide useful information. The comprehensive saving of results and model artifacts allows for further validation and deployment.

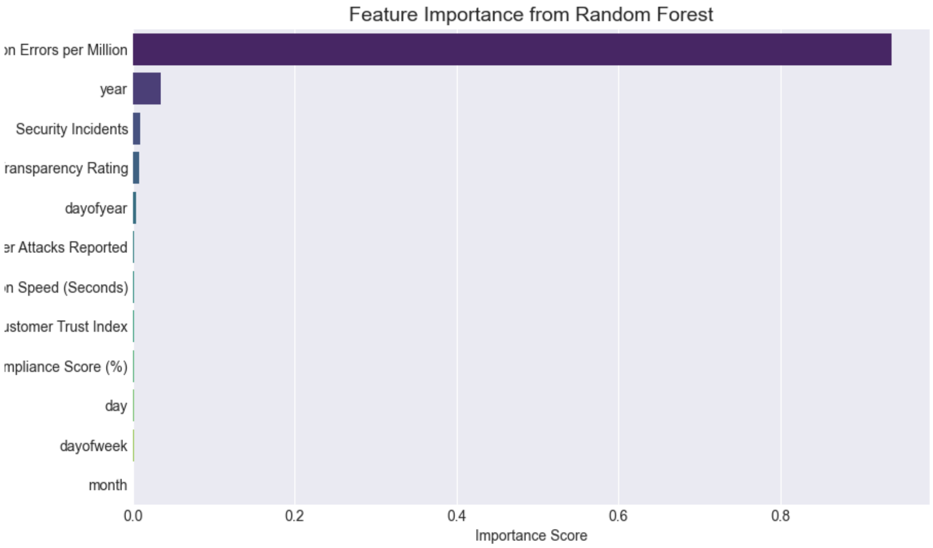

Figure 6 Feature Importance Plots

Figure 6 presents the feature importance scores derived from the Random Forest model. This horizontal bar chart ranks all input features based on their contribution to predicting fraud cases. The overwhelming majority of predictive power is attributed to the feature “Transaction Errors per Million,” which dominates the plot with an importance score close to 0.94. This confirms that transaction errors are the primary driver in the model’s fraud prediction capability. Other features such as “year” and “Security Incidents” have significantly smaller importance scores, highlighting their secondary roles. The visualization clearly demonstrates that while several features were included, the model relies heavily on transactional error rates to detect fraud patterns, emphasizing its critical role in fraud monitoring systems.

Discussion

The Random Forest Regression model demonstrated a strong ability to accurately predict fraud cases in e-commerce transactions, as evidenced by the high R-squared value and low error metrics. This suggests that the model can effectively capture the complex relationships between transaction errors, security incidents, and other relevant features with fraud occurrences. Such predictive power is valuable for developing more responsive and data-driven fraud detection systems, enabling businesses to anticipate fraudulent activity and take preventive measures in a timely manner. However, despite these promising results, the model has several limitations that must be acknowledged. One key challenge is handling potential imbalances in the dataset, as fraud cases and other related events may not be evenly distributed across time or transaction types. Imbalanced data can cause the model to underperform on rare but critical fraud instances, which may require additional techniques such as data augmentation or specialized algorithms to address. Moreover, the model’s complexity and tuning parameters, like maximum tree depth and minimum samples split, must be carefully managed to avoid overfitting, where the model performs well on training data but poorly on unseen data. Finally, while the Random Forest model provides insights through feature importance, it remains a black-box approach in terms of interpretability compared to simpler models. This can pose challenges when explaining fraud detection decisions to stakeholders or regulators, which is an important consideration in the context of cyberlaw and compliance. Future work should explore methods to improve model transparency and robustness, as well as testing the model on more diverse datasets to ensure its generalizability in real-world fraud detection applications.

Conclusion

This study demonstrated that Random Forest Regression is an effective method for predicting fraud cases in e-commerce transactions. The model showed high accuracy and was able to capture key relationships between transaction-related variables and fraud occurrence. By enabling more accurate forecasting of fraudulent activity, this approach can help improve transaction security and reduce losses in online marketplaces. The research contributes to the expanding field of applying machine learning techniques to cybersecurity challenges, particularly fraud detection. It provides empirical evidence that advanced data mining methods can support proactive fraud prevention strategies. This work adds value to both academic research and practical implementations, offering a scalable solution that can be adapted and enhanced for real-world use. Looking ahead, future research could focus on enhancing model performance by incorporating additional relevant features or experimenting with alternative machine learning algorithms to improve robustness and interpretability. Furthermore, the findings have important implications for cyberlaw and regulatory frameworks, highlighting how sophisticated fraud detection tools can support enforcement efforts and promote safer e-commerce environments through better monitoring and compliance mechanisms.

Declarations

Author Contributions

Conceptualization: Y.D.; Methodology: Y.D.; Software: A.W.B.A.R.; Validation: A.W.B.A.R.; Formal Analysis: A.W.B.A.R.; Investigation: Y.D.; Resources: A.W.B.A.R.; Data Curation: A.W.B.A.R.; Writing Original Draft Preparation: Y.D.; Writing Review and Editing: Y.D.; Visualization: Y.D.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] R. Damayanti and Z. Adrianto, “Machine Learning for E-Commerce Fraud Detection,” *J. Ris. Akunt. Dan Bisnis Airlangga*, vol. 8, no. 2, pp. 1562-1577, 2023, doi: 10.20473/jraba.v8i2.48559.

[2] A. Mutemi and F. Bação, “E-Commerce Fraud Detection Based on Machine Learning Techniques: Systematic Literature Review,” *Big Data Min. Anal.*, vol. 7, no. 2, pp. 419-444, 2024, doi: 10.26599/bdma.2023.9020023.

[3] C. S. Lee, “How Online Fraud Victims Are Targeted in China: A Crime Script Analysis of Baidu Tieba C2C Fraud,” *Crime Delinquency*, vol. 68, no. 13, pp. 2529-2553, 2021, doi: 10.1177/00111287211029862.

[4] N. A. Hidayah Razali, W. R. Wan Rosli, and M. B. Othman, “The Legal Protection of E-Consumers Against E-Commerce Fraud in Malaysia,” *Malays. J. Soc. Sci. Humanit. Mjssh*, vol. 7, no. 9, p. e001778, 2022, doi: 10.47405/mjssh.v7i9.1778.

[5] J. Li, “E-Commerce Fraud Detection Model by Computer Artificial Intelligence Data Mining,” *Comput. Intell. Neurosci.*, vol. 2022, no. 5, pp. 1–9, 2022, doi: 10.1155/2022/8783783.

[6] M. Misirana *et al.*, “Early Detection Method for Money Fraudulent Activities on E-Commerce Platform via Sentiment Analysis,” *J. Entrep. Bus.*, vol. 9, no. 2, pp. 121–142, 2021, doi: 10.17687/jeb.v9i2.804.

[7] H. Du, D. Li, and W. Wang, “Abnormal User Detection via Multiview Graph Clustering in the Mobile E-Commerce Network,” *Wirel. Commun. Mob. Comput.*, 2022, vol. 2022, no. 8, p. e3766810, doi: 10.1155/2022/3766810.

[8] Y. Wang, W. Yu, P. Teng, G. Liu, and D. Xiang, “A Detection Method for Abnormal Transactions in E-Commerce Based on Extended Data Flow Conformance Checking,” *Wirel. Commun. Mob. Comput.*, vol. 2022, no. 1, p. e4434714, 2022, doi: 10.1155/2022/4434714.

[9] M. D. Venkata Prasad, “Multi-Entity Real-Time Fraud Detection System Using Machine Learning: Improving Fraud Detection Efficiency Using FROST-Enhanced Oversampling,” *Jes*, 2024, vol. 20, no. 7, pp. 1380-1394, doi: 10.52783/jes.3710.

[10] S. Rout and K. L. Jaiswal, “Fraud Detection Using Deep Learning,” *Int. J. Electr. Data Commun.*, vol. 5, no. 1, pp. 7-11, 2024, doi: 10.22271/27083969.2024.v5.i1a.37.

[11] L. S. Hasugian and S. Suharjito, “Fraud Detection for Online Interbank Transaction Using Deep Learning,” *Syntax Lit. J. Ilm. Indones.*, vol. 8, no. 6, pp. 4263-4275, 2023, doi: 10.36418/syntax-literate.v8i6.12627.

[12] S. F. Pratama and A. M. Wahid, “Fraudulent Transaction Detection in Online Systems Using Random Forest and Gradient Boosting,” *J. Cyber Law*, vol. 1, no. 1, pp. 88-115, Mar. 2025, doi: https://doi.org/10.63913/jcl.v1i1.5.

[13] A. Taha, “Intelligent Ensemble Learning Approach for Phishing Website Detection Based on Weighted Soft Voting,” *Mathematics*, vol. 9, no. 21, p. 2799, 2021, doi: 10.3390/math9212799.

[14] X. Li, W. Zhang, D. Wang, B. Zhang, and H. He, “Algorithm of Web Page Similarity Comparison Based on Visual Block,” *Comput. Sci. Inf. Syst.*, vol. 16, no. 3, pp. 815-830, 2019, doi: 10.2298/csis180915028l.

[15] R. Gupta, “Data Mining for Fraud Detection: An Overview of Techniques and Applications,” *Turk. J. Comput. Math. Educ. Turcomat*, vol. 10, no. 1, pp. 561-567, 2019, doi: 10.17762/turcomat.v10i1.13549.

[16] A. A. A. Mousa, “Detecting Financial Fraud Using Data Mining Techniques: A Decade Review From 2004 to 2015,” *J. Data Sci.*, vol. 14, no. 3, pp. 553-570, 2022, doi: 10.6339/jds.201607_14(3).0010.

[17] A. R. Dastjerdi, D. Foroghi, and G. H. Kiani, “Detecting Manager’s Fraud Risk Using Text Analysis: Evidence From Iran,” *J. Appl. Account. Res.*, vol. 20, no. 2, pp. 154-171, 2019, doi: 10.1108/jaar-01-2018-0016.

[18] S. Cho, “Fraud Detection in Malaysian Financial Institutions Using Data Mining

and Machine Learning," *J. Inf. Technol.*, vol. 7, no. 1, pp. 13-21, 2023, doi: 10.53819/81018102t4152.

[19] S. Meghana and S. K. R, "An Efficient Approach to Detect Fraudulent Service Enrollment Websites With Novel Random Forest and Compare the Accuracy With XGBoost Machine Algorithm," *E3s Web Conf.*, vol. 399, no. 4, p. 04022, 2023, doi: 10.1051/e3sconf/202339904022.

[20] R. Pk, "Enhanced Credit Card Fraud Detection: A Novel Approach Integrating Bayesian Optimized Random Forest Classifier With Advanced Feature Analysis and Real-Time Data Adaptation," *Int. J. Innov. Eng. Manag. Res.*, vol. 12, no. 5, pp. 537-561, 2023, doi: 10.48047/ijiemr/v12/issue05/52.

[21] M. Lokanan, "Predicting Mobile Money Transaction Fraud Using Machine Learning Algorithms," *Appl. Ai Lett.*, vol. 4, no. 2, pp. 1-10, 2023, doi: 10.1002/ail2.85.

[22] Y. Lucas *et al.*, "Towards Automated Feature Engineering for Credit Card Fraud Detection Using Multi-Perspective HMMs," *Future Gener. Comput. Syst.*, vol. 102, no. 1, pp. 393-402, 2020, doi: 10.1016/j.future.2019.08.029.

[23] H. Xu, G. Fan, and Y. Song, "Novel Key Indicators Selection Method of Financial Fraud Prediction Model Based on Machine Learning Hybrid Mode," *Mob. Inf. Syst.*, vol. 2022, no. 3, pp. 1-10, 2022, doi: 10.1155/2022/6542652.