Fraud detection in E-Commerce Transactions

Kevin Vagan Djafar#1, Micko Melvern\*2,William Raffa Dharmawan#3

***Abstract*— This paper uses two integrated datasets to present a machine learning-based approach for fraud detection in e-commerce transactions. The study applies comprehensive data preprocessing techniques, including feature transformation (e.g., converting IP addresses to numeric values, calculating account age) and imputing missing values through probabilistic sampling. Key transaction-related features such as customer demographics, payment method, device used, and browser type are analysed to identify patterns indicative of fraudulent behaviour. Statistical methods are employed to compute frequencies and percentages of categorical variables, providing insights into feature distributions. The findings demonstrate how preprocessing and feature engineering improve model performance, enabling more accurate identification of fraudulent transactions in e-commerce platforms.**

***Keywords***— **Fraud detection, e-commerce, machine learning, data preprocessing, feature engineering.**

1. Introduction

Online retailers are increasingly concerned about fraudulent activities in e-commerce, which can lead to significant financial losses and harm to a brand's reputation. Traditional methods of detecting fraud are frequently inadequate to meet the complexity and scale of today's e-commerce platforms, especially as the volume of online transactions continues to increase. This has prompted the use of machine learning methods, which present viable ways to automate and raise the degree of accuracy of fraud detection systems.

Global e-commerce fraud is rapidly escalating, with losses projected to exceed $48 billion in 2023, and cumulative losses expected to surpass $343 billion by 2027 (Mastercard, 2024). Account takeover fraud rose to 29% of all fraud in 2023, and chargeback fraud accounted for 34% of global eCommerce fraud, while losses from triangulation fraud are estimated between $660 million and $1 billion monthly (WiserNotify, n.d.). To mitigate this, 75% of eCommerce businesses plan to increase their fraud prevention budgets, while leveraging tools like two-factor authentication and machine learning to combat rising threats such as phishing, account takeovers, and return fraud (Howarth, 2024).

The rise in fraud has also led to increased attention to the role of advanced data science techniques in fraud prevention. Machine learning models offer the ability to analyse vast amounts of data in real time, identifying anomalies and suspicious patterns that would be difficult for traditional systems to detect. Furthermore, integrating machine learning with other technological advancements, such as natural language processing and behaviour analytics, enhances predicting and preventing fraud (Guo et al., 2018). For example, analysing patterns in user behaviour—like login attempts, browsing habits, and transaction timing—can help systems flag potential fraudulent activities with greater precision. Advanced algorithms can also adapt to evolving fraud tactics, making them crucial for keeping pace with increasingly sophisticated fraud schemes.

In this paper, we propose a framework for detecting fraud in e-commerce transactions by combining machine learning algorithms with sophisticated data preprocessing techniques. In particular, we prioritise integrating various datasets, transforming important features (like IP addresses and account age), and using probabilistic imputation to handle missing values. We want to find patterns that can successfully differentiate between authentic and fraudulent transactions by examining features like customer demographics, transaction details, and device/browser information. By showing how feature engineering and preprocessing can greatly improve the performance of machine learning models in an e-commerce setting, our work adds to the developing field of fraud detection.

1. Methodology

To develop our model, we have decided to implement data mining models to help in detecting fraud based on our dataset.

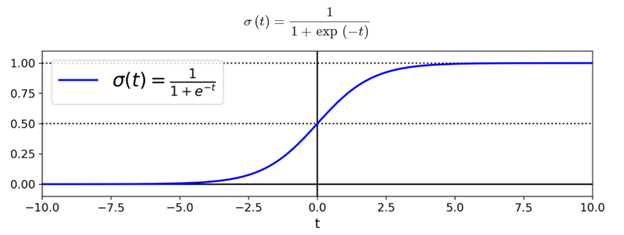
This section outlines the methodologies and theoretical foundations employed in the study of fraud detection in e-commerce transactions. The methods include logistic regression (LogReg), Random Forest (RF), and Extreme Gradient Boosting (XGBoost). The choice of these algorithms was driven by their suitability for classification tasks and their ability to model complex data relationships effectively.

1. *Pre-Processing*

Data preprocessing is a critical step to ensure the quality and reliability of the input data. The following steps were applied:

* Data Cleaning: Missing values were checked, and invalid entries, such as negative customer ages, were filtered out to ensure data integrity.
* Feature Engineering: Relevant features such as transaction day, month, and day of the week were extracted from date fields. Categorical variables were one-hot encoded, and numerical features were standardised to improve algorithm performance.
* Balancing the Dataset: The dataset, which was initially imbalanced, was balanced using the ROSE (Random Over-Sampling Examples) package to ensure equal representation of fraudulent and non-fraudulent transactions. This step mitigates bias in model predictions.
* Data Splitting: The dataset was divided into training, validation, and test sets in a 70:15:15 ratio to prevent overfitting and evaluate model generalisation.

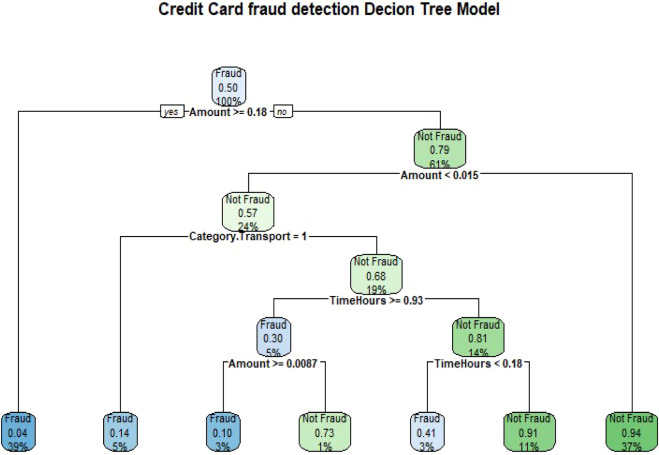
1. *Logistic Regression (LogReg)*

**

Logistic regression is a statistical method used for binary classification problems. It models the probability of an event occurring as a function of the input features. The model assumes a linear relationship between the independent variables and the log odds of the dependent variable (Maalouf, 2011). The probability function is given as where it is the probability of the positive class, is the intercept, are the coefficients and are the independent variables.

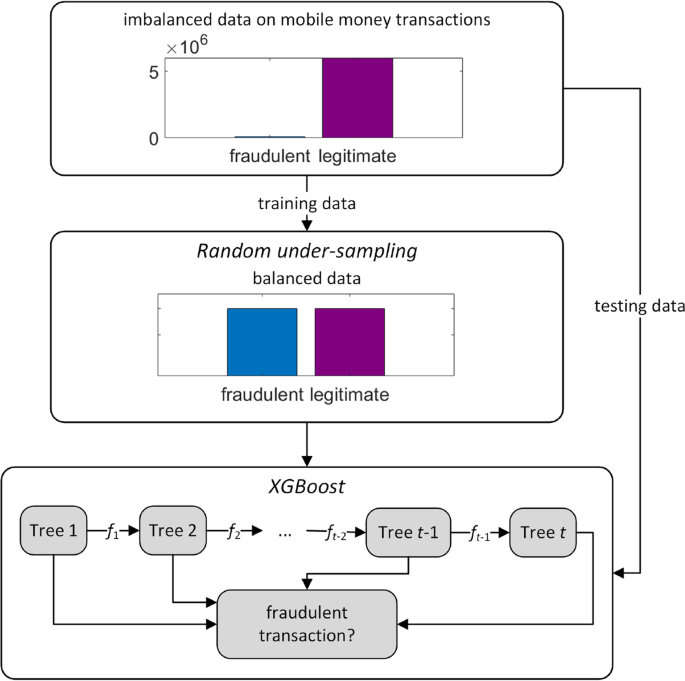
The model was optimised using maximum likelihood estimation and evaluated using metrics such as accuracy, precision, recall, and F1-score. The model evaluation also included generating a ROC curve and calculating the AUC.

1. *Random Forest (RF)*

**

Random Forest is an ensemble learning method that constructs multiple decision trees during training and aggregates their outputs. Each tree is trained on a bootstrap dataset sample, while a random subset of features is considered for splitting at each node. Predictions are made via majority voting for classification tasks, which reduces overfitting and increases robustness. The mathematical foundations include Gini impurity for measuring node purity and information gain for evaluating the effectiveness of splits. The Random Forest model was trained with 50 trees and evaluated using confusion matrices, accuracy, precision, recall, F1-score, and AUC (Ziegler & König, 2013).

1. *Extreme Gradient Boosting (XGBoost)*

**

XGBoost is a gradient-boosting algorithm that builds trees sequentially, with each new tree correcting the errors of its predecessors. The algorithm combines a loss function and a regularization term in its objective function, which helps in preventing overfitting and optimizing performance. It employs second-order Taylor expansion to approximate the loss function, leveraging the gradient and Hessian for efficient optimisation (Yulianti et al., 2022). The XGBoost model was trained with 100 rounds using an AUC-based evaluation metric. Predictions were thresholded at 0.7 for binary classification, and their performance was assessed using metrics such as accuracy, precision, recall, F1-score, and AUC.

1. *Modelling*

Three machine learning models were employed: Logistic Regression, Random Forest, and XGBoost. Logistic Regression served as a baseline model for its simplicity and interpretability. Random Forest offered robust performance through ensemble learning and feature importance analysis. XGBoost provided an advanced gradient-boosting approach, balancing accuracy and computational efficiency (Zheng et al., 2024).

1. *Methodological Approach*

The methodology follows the Knowledge Discovery in Databases (KDD) process, which includes several systematic stages:

* Data Collection

The dataset was sourced from Kaggle and comprises transaction attributes such as amount, customer age, account age, and transactional time. The target variable indicates whether a transaction is fraudulent.

* Data Preprocessing

To ensure data quality, missing values were handled, and invalid entries, such as negative customer ages, were removed. Features like transaction day, month, and day of the week were extracted. Categorical variables were one-hot encoded, and numerical features were standardized. To address the class imbalance, the dataset was balanced using the ROSE (Random Over-Sampling Examples) method. The data generated using the ROSE method provides a more accurate representation of the original dataset.

* Modelling

Three machine learning models were employed: Logistic Regression, Random Forest, and XGBoost. Logistic Regression served as a baseline model for its simplicity and interpretability. Random Forest offered robust performance through ensemble learning and feature importance analysis. XGBoost provided an advanced gradient-boosting approach, balancing accuracy and computational efficiency.

* Model Evaluation

The models were evaluated using metrics including accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC-ROC). These metrics provided a comprehensive understanding of model performance and highlighted trade-offs between precision and recall.

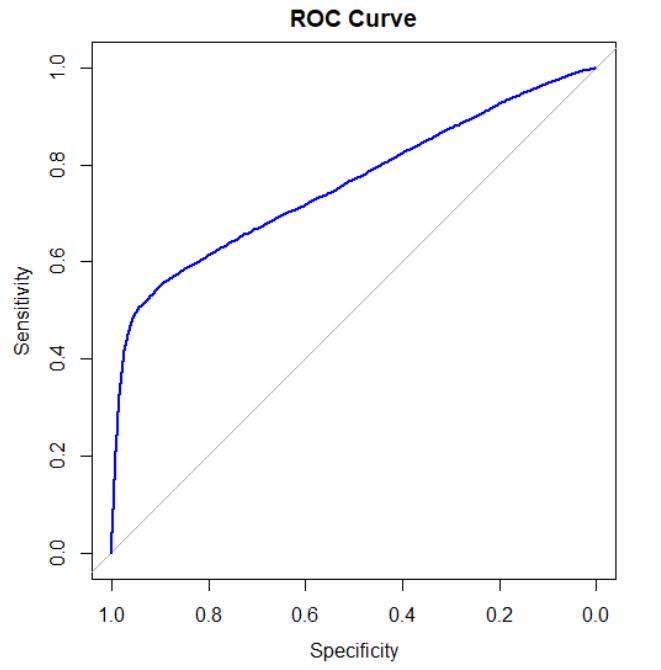
1. Result
2. Metrics used to measure

The code and the model are accessible through this Google Drive link: <https://drive.google.com/file/d/1Ms2D-XHayprI6pgdMo35xXSLSqMopYD3/view?usp=sharing>, which are used to determine the results of each model.

The models will be compared using confusion matrix, accuracy, precision, recall, F1 score, ROC curve and AUC. A confusion matrix is used to show a comparison between predicted classification and actual classification. The accuracy score will show the ratio of correctly predicted instances to the total prediction amount. precision will show the correctly predicted positives out of all the predicted positives. recall will show the ratio of correctly predicted positives to all actual positives. F1 will be the most useful for these metrics, balancing between precision and recall and usually the metrics to use in fraud detection.

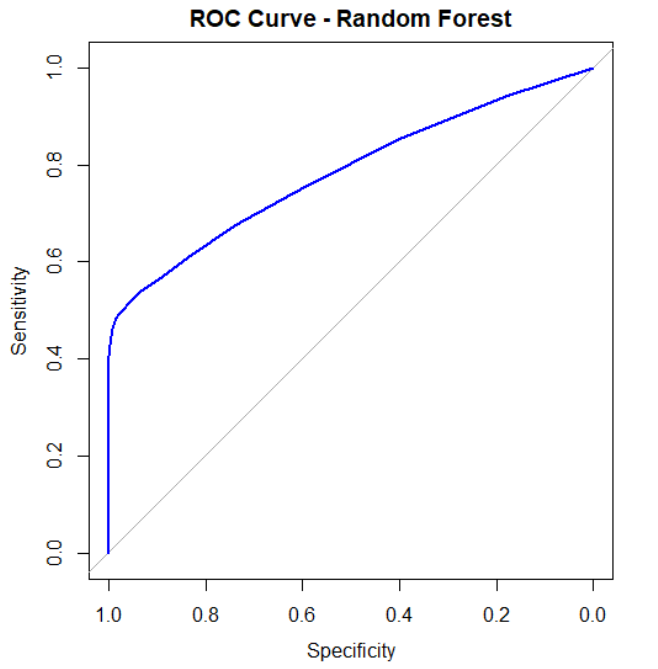
1. Without ROSE
2. Logistic Regression

|  | Actual | |
| --- | --- | --- |
| Predicted | 0 | 1 |
| 0 | 83493 | 6394 |
| 1 | 1 | 104 |



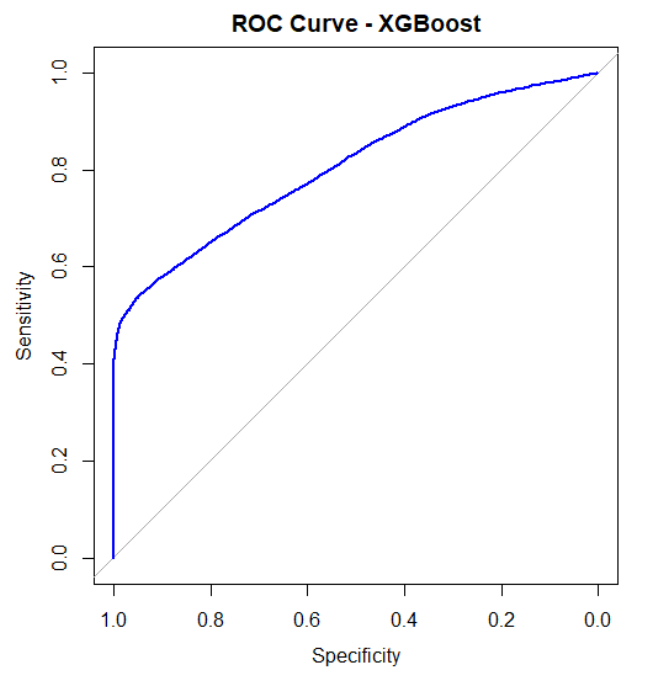
1. Random Forest

|  | Actual | |
| --- | --- | --- |
| Predicted | 0 | 1 |
| 0 | 83395 | 3865 |
| 1 | 99 | 2633 |



1. XGBoost

|  | Actual | |
| --- | --- | --- |
| Predicted | 0 | 1 |
| 0 | 83443 | 3920 |
| 1 | 51 | 2578 |



Comparison Table

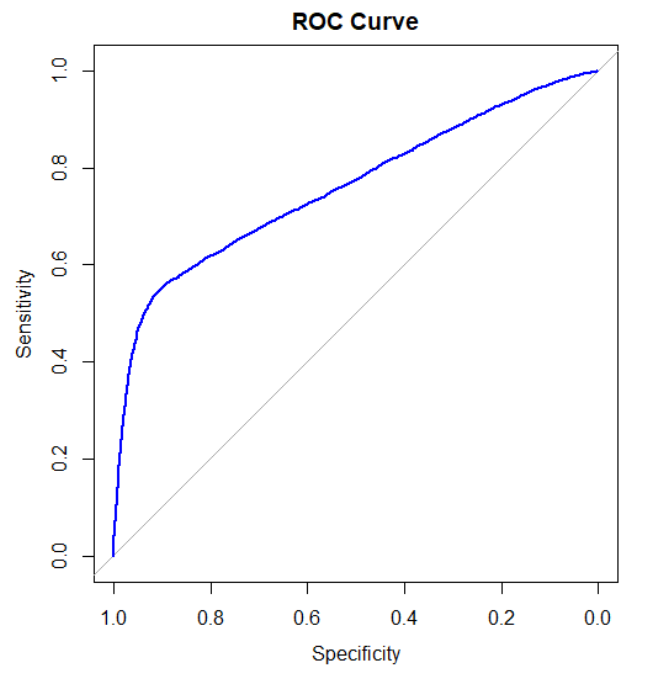
|  | Log Reg | Random Forest | XGBoost |
| --- | --- | --- | --- |
| Accuracy | 92.89% | 95.6% | 95.59% |
| Precision | 99.05% | 96.38% | 98.06% |
| Recall | 1.6% | 40.52% | 39.67% |
| F1 Score | 3.15% | 57.05% | 56.49% |
| AUC | 0.76 | 0.78 | 0.8 |

The table above compares the performance of the three models, Random Forest offers the best accuracy out of the models (95.6%), while also giving the best results for recall (40.52%) and F1 score (57.05%), it is the best model for balanced performance across all metrics. Logistic regression offers the best precision (99.05%) out of all the models, and shows a competent accuracy score (92.89%), but it did not beat the other models, logistic regression performed very poorly in the recall and F1 score, only getting 1.6% and 3.15% respectively. XGBoost shows a similar performance to random forest, but it has the highest AUC (0.8) out of all the model

C. With ROSE

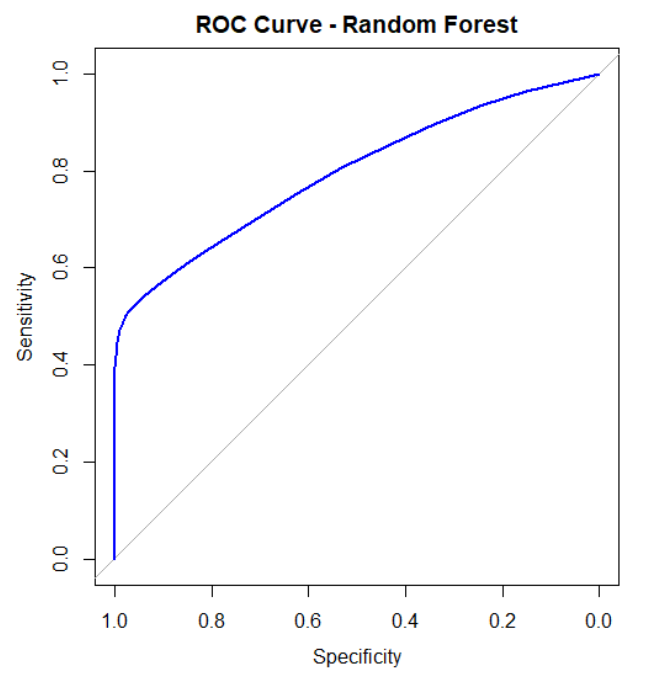
1. Logistic Regression

|  | Actual | |
| --- | --- | --- |
| Predicted | 0 | 1 |
| 0 | 78987 | 3416 |
| 1 | 4507 | 3082 |



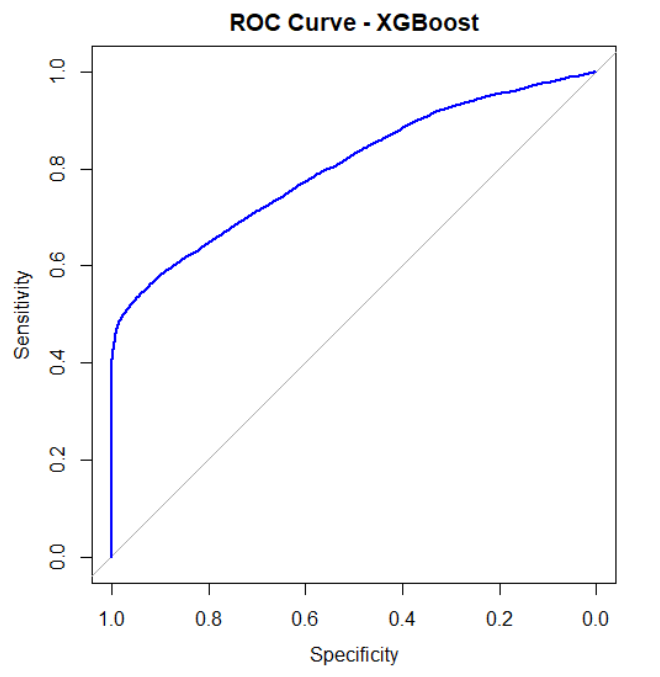
1. Random Forest

|  | Actual | |
| --- | --- | --- |
| Predicted | 0 | 1 |
| 0 | 82769 | 3508 |
| 1 | 725 | 2990 |



1. XGBoost

|  | Actual | |
| --- | --- | --- |
| Predicted | 0 | 1 |
| 0 | 82032 | 3330 |
| 1 | 1462 | 3168 |



Comparison Table

|  | Log Reg | Random Forest | XGBoost |
| --- | --- | --- | --- |
| Accuracy | 91.20% | 95.3% | 94.68% |
| Precision | 40.61% | 80.48% | 68.42% |
| Recall | 47.43% | 46.01% | 48.75% |
| F1 Score | 43.76% | 58.55% | 56.94% |
| AUC | 0.76 | 0.8 | 0.8 |

The table above compares the performance of the three models using ROSE. ROSE generally improves the recall and F1 score of the model that it's used with, this can be seen in the result of logistic regression, the recall score improves significantly from 1.6% to 47.43% and the F1 score improves from 3.15% to 43.76%, but using ROSE also shows a major decrease in the precision score, from 99.05% to 40.61%, this trend can be seen in all models, random forest precision is reduced from 96.38% to 80.48% and XGBoost precision is reduced from 98.06% to 68.42%. Using ROSE also slightly decreases the accuracy, around 1% average accuracy is lost between the three models. This trend also follows for the recall and F1 score in random forest and XGBoost. ROSE increases the AUC in random forest from 0.78 to 0.8.

1. Conclusions

The comparison between the three models shows distinct strengths and weaknesses across different evaluation metrics. Random forest is the most balanced out of the three models, offering the highest accuracy (95.6%), recall (40.52%), and F1 score (57.05%), making it the best model overall. Logistic regression reached the highest precision (99.05%) out of all the models but performed very poorly in recall (1.6%) and F1 score (3.15%), XGBoost performed very similarly to the random forest but achieved the highest AUC (0.8). The application of ROSE improved the F1 score of the logistic regression model(1.6% to 47.43%) and recall (3.15% to 43.76%), albeit with a significant drop in precision (99.05% to 40.61%). ROSE is most suited for logistic regression, the other model almost does not get any performance increase in F1 score and Recall (around 1% each), but still, drops their precision from 96.38% to 80.48% in random forest and from 98.06 to 68.42% in XGBoost

1. References

Guo, C., Wang, H., Dai, H.-N., Cheng, S., & Wang, T. (2018, October 28). *Fraud Risk Monitoring System for E-Banking Transactions*. 10.1109/DASC/PiCom/DataCom/CyberSciTec.2018.00030

Howarth, J. (2024, May 15). *23+ eCommerce Fraud Statistics (2024)*. Exploding Topics. https://explodingtopics.com/blog/ecommerce-fraud-stats

Maalouf, M. (2011, July). Logistic regression in data analysis: An overview. *International Journal of Data Analysis Techniques and Strategies*, *3*, 281-299. 10.1504/IJDATS.2011.041335

Mastercard. (2024). *Ecommerce fraud trends and statistics merchants need to know*. Mastercard. https://b2b.mastercard.com/news-and-insights/blog/ecommerce-fraud-trends-and-statistics-merchants-need-to-know-in-2024/

WiserNotify. (n.d.). *E-Commerce Fraud Statistics: $48 Billion Lost Annually*. https://wisernotify.com/blog/ecommerce-fraud-stats/

Yulianti, S. E. H., Sukmawaty, Y., & Soesanto, O. (2022). Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit. *Journal of Mathematics Theory and Application*, *4*, 21-26. 10.31605/jomta.v4i1.1792

Zheng, Q., Yu, C., Cao, J., Xu, Y., Xing, Q., & Jin, Y. (2024, June 7). Advanced Payment Security System:XGBoost, LightGBM and SMOTE Integrated. http://arxiv.org/abs/2406.04658

Ziegler, A., & König, I. (2013, December 23). Mining data with random forests: Current options for real-world applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *4*. 10.1002/widm.1114