ASIAN JOURNAL OF MATHEMATICAL SCIENCES

AJMS

Asian Journal of Mathematical Sciences

RESEARCH ARTICLE

Detecting Fraud Transactions in Financial Institutions

Awogbemi Clement Adeyeye1* Dayo, Kayode Vincent2 Ilori, Adetunji Kolawole1 Oyeyemi, Gafar Oyeyemi3

1Statistics Programme, National Mathematical Centre, Abuja, Nigeria
2 Statistics Department, University of Abuja, Abuja, Nigeria
3 Statistics Department, University of Ilorin, Ilorin, NIgeria

Corresponding Email: awogbermiadeyeye@yahoo.com

Received: 19-07-2025; Revised: 05-08-2025; Accepted: 11-09-2025

ABSTRACT

Detecting fraud and anomalies in financial transactions is crucial in safeguarding institutional assets, maintaining regulatory compliance and ensuring customers trust in financial system. This study investigated methods of detecting frauds or anomalies in transactions within financial institutions, a vital task to prevent financial losses, reduce investigative costs, and comply with regulatory standards. We compared the efficiency of three statistical models: Logistic Regression, Linear Discriminant analysis (LDA).and Quadratic Discriminant (QDA), in identifying fraudulent activity. Secondary data of over 280,000 financial transactions from an online website (Kaggle) was used to evaluate each model based on accuracy, precision, and error rates, for both fraudulent and non-fraudulent classifications. The results indicated that Logistic Regression outperformed LDA, and QDA, achieving the highest accuracy and lowest error rate, making it the most effective model among the models considered in the study for fraud detection in this context.

Keywords: Fraud Transactions, Anomalies, Discriminant Analysis, Financial Institutions, Logistic Regression.

INTRODUCTION

Electronic banking services have faced regulatory pressures to secure it operations while out sourcing critical infrastructure like ATM to private operator. This threat has highlighted the urgent need to effectively detect fraud mechanisms and hence safeguard financial transactions and maintain the integrity of electronic banking services (Kian, 2022).

Risk management in financial institutions relies heavily on risk models to quantify and control anomalies or outliers in financial transactions. The insights and timely risk metrics provided by these essential models do not only inform decision making, but also help regulatory standards (Crépey et al., 2022).

Financial institutions around the world suffer impacts of fraudulent monetary transactions. Protection measures for either face-to-face or online banking transactions suffer some sort of scam thereby subjecting vulnerable bank users to financial fraud impact or it equivalences (Torres & Ladeira, 2020). Managing these grievous challenges using multivariate statistical models and as well compare their performances in correctly classifying fraudulent financial transactions is the focus of this paper. Financial fraud has garnered much more attention in the past decades due to the potential consequences of undetected anomalies within the industry and our everyday life. These crimes can vary in nature and have the effect of possibly destabilizing economies, increasing the cost of living and impacting the consumer's sense of security (Hilal et al., 2022).

Agresti (2002) extended the ordinary regression concept to logistic regression models to include qualitative multiple explanatory variables, often called factors. He defined the model

$\pi(X)=p(Y=1)$ at values $x= (x_1, x_2, x_p)$ of p predictors as

$$\text{logit}[\pi(X)] = \propto + \beta_1 x_1+\beta_2 x_2+ \ldots+\beta_p x_p \qquad (1)$$

The parameter β_i refers to the effect of x_i on the log odds the Y= 1, controlling the other x_j. For instance, $\exp(\beta_i)$ is the multiplicative effect on the odds of a 1-unit increase in x_i, at fixed level of other x_j. An explanatory variable can be qualitative, using dummy values for categories.

Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) are valuable tools in detecting fraud within financial transactions by classifying transaction patterns.LDA is effective when classes share similar covariances, while QDA accommodates differing covariances, aiding in distinguishing complex transaction behaviours. These techniques improve precision in identifying typical versus suspicious transactions, strengthening fraud detection efforts (Bolton & Hand, 2002). Therefore, study compares the performance of the three models in determining the factors that best describe fraudulent financial transactions in our financial institutions.

METHODOLOGY

Logistic Regression as explained by (Scott et al., 1991) is used for binary classification, modelling the probability $P(Y=1/X)$, of an outcomes $Y=1$, given a vector of predictors $X= X_1, X_2\ldots, X_k$. The values take the form

$$P(Y=1/X) =1/ (1+e^{(-(A_0+A_1 X_1+A_2 X_2+ \ldots+A_k X_k))}) \qquad (2)$$

where

Y is the binary dependent variable (0 or 1)

x_i, i=1, 2, k are the predictor variables

A_i i=0,1, 2…, k are the coefficients of the predictor estimated by maximizing the likelihood function.

Linear Discriminant Analysis (LDA) aims to find a linear combination of features that best separates two or more classes. Assuming classes have a common covariance matrix, LDA classifies based on the following decision rule:

$$\delta_k (X)=X^{'} \sum^{(-1)} \mu_k-1/2\, \mu_k^{'} \sum^{(-1)} \mu_k+ \text{lin}\, (\pi_k) \qquad (3)$$

where μ_k is the mean vector of the class k, $\sum$ is the shared covariance matrix and π_k is the prior probability of class k (Hastie, T., Tibshirani, R., & Friedman, 2009).

Quadratic Discriminant Analysis (QDA) generalizes LDA by allowing each class to have it own covariance matrix, offering more flexibility. The decision rule is

$$\delta_k (X)=1/2\, \text{lin}|\sum_k | -1/2\, (X - \mu_k)^{'} \sum_k^{(-1)} (X- \mu_k) + \text{lin}(\pi_k) \qquad (4)$$

where $\sum_k$ is the covariance matrix specific to class k, and other terms are as in LDA. QDA is advantageous when the assumption of equal covariance does not hold (James et al., 2013).

R (caTools, ROCR, tidy verse, caret) programming was used to fit the three models (Logistic Regression, Linear and Quadratic Discriminant models).

Data Description

The data used are secondary data obtained online from Kaggle with the following features: total observations =284807, Amount of Transaction, Class (1= Fraud, 0 = no fraud). The 28 other variables (V1, V2, …, V27, V28). were labelled, but not the real representation was kept hidden for confidential reasons.

Presentation of Tables

The classification error rate is the proportion of incorrect predictions over the total number of predictions. Given false negatives (FN), false positives (FP), true positives (TP), and true negatives (TN), the error can be as expressed by (Powers, 2020)

Error Rate=(FP+FN)/ (TP+TN+FP+FN) (5)

Table 1: The confusion matrix from the Logistic model

	Predicted (0)	**Predicted (1)**	**Error**
Actual (0)	284273	42	0.000794
Actual (1)	184	308	

Table 2: Confusion matrix for the LDA

	Predicted (0)	**Predicted (1)**	**Error**
Actual (0)	284256	115	0.000611
Actual (1)	59	377	

Table 3: Confusion matrix for the QDA

	Predicted (0)	**Predicted (1)**	**Error**
Actual (0)	277263	60	0.024971
Actual (1)	7052	432	

RESULT AND DISCUSSION

Logistic Regression (Table 1) performed the best comparing the false positives (42). It has the lowest classification error for both classes, especially with high accuracy in predicting class (0) and reasonably good performance in predicting class (1), making it the overall best model for the predictors with error rate of 0.000794.

The LDA model (Table 2) performed better than the QDA with the lowest error rate (0.000611), but slightly worse than logistic regression (False positive = 115). It has a slightly high error rate compared to logistic regression but yet captured both classes, especially with fewer false negatives for class (1).

The QDA model (Table 3) performed the poorest. Although it has a smaller false positive of 60, but it is significantly higher error rate (0.024971), particularly with a very high false negative count for class (1), showing that it struggled to accurately classified the minority that both the logistic regression and the LDA models.

CONCLUSION

In this study, three distinct models (Logistic Regression, Linear Discriminant Analysis, and Quadratic Discriminant Analysis) were explored to detect fraudulent transactions within financial institutions. Based on the results, Logistic Regression demonstrated better performance, having the highest accuracy and lowest error rate among the three models, LDA performed moderately well bur QDA exhibited significantly higher error rate, especially in the misclassification of fraudulent transactions. The findings show the potential of Logistic Regression as a reliable tool for fraud detection, given it balanced approach in managing false positives and false negatives effectively.

RECOMMENDATION

To enhance fraud detection accuracy in financial institutions, we recommend prioritizing Logistic Regression models due to their robust performance across different transaction types. Furthermore, financial institutions should consider continuous updating model parameters to account for evolving fraud patterns and to improve detection rates. Additional research could explore the interaction of ensemble models or hybrid approaches that combine the strengths of different classification techniques for more comprehensive fraud detection system.

REFERENCES

1. Agresti, A. (2002). Categorical Data Analysis. In A John Wiley & Sons INC, (Issue 4).

2. Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. Statistical Science, 17(3), 235–255. https://doi.org/10.1214/ss/1042727940.

3. Crépey, S., Lehdili, N., Madhar, N., & Thomas, M. (2022). Anomaly Detection in Financial Time Series by Principal Component Analysis and Neural Networks. Algorithms, 15(10), 1–38. https://doi.org/10.3390/a15100385

4. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. In Journal of the American Geriatrics Society (Vol. 32, Issue 6). https://doi.org/10.1111/j.1532-5415.1984.tb02220.x

5. Hilal, W., Gadsden, S. A., & Yawney, J. (2022). Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances. Expert Systems with Applications, 193, 116429. https://doi.org/10.1016/j.eswa.2021.116429.

6. James, G., Witten, D., & Hastie, T. (2013). An Introduction to Statistical Learning (1st Edition). Springer.

7. Kian, R. (2022). Journal of Applied Research on Industrial Engineering Paper Type: Research Paper Detection of Fraud in Banking Transactions Using Big Data Clustering Technique Customer Behavior Indicators 1 | Introduction 2 | Literature Review. 9(3), 264–273.

8. Powers, D. M. W. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. 37–63. http://arxiv.org/abs/2010.16061

9. Scott, A. J., Hosmer, D. W., & Lemeshow, S. (1991). Applied Logistic Regression. Biometrics, 47(4), 1632. https://doi.org/10.2307/2532419

10. Torres, R. A. L., & Ladeira, M. (2020). A proposal for online analysis and identification of fraudulent financial transactions. Proceedings - 19th IEEE International Conference on Machine Learning and Applications, ICMLA 2020, 240–245. https://doi.org/10.1109/ICMLA51294.2020.00047