ENHANCED KNOWLEDGE IN SCIENCES AND TECHNOLOGY
e-ISSN: 2773-6385

EKST

Vol. 4 No. 2 (2024) 201-211
https://publisher.uthm.edu.my/periodicals/index.php/ekst

# LASSO and Elastic Net on Bank Account Fraud Detection

**Tan Jing Thong[1], Khuneswari Gopal Pillay[1]***

[1] *Department of Mathematics and Statistics, Faculty of* Applied *Sciences and Technology, UTHM Kampus Cawangan Pagoh, Hab Pendidikan Tinggi Pagoh, KM 1, Jalan Panchor, 84600 Pagoh, Muar, Johor, MALAYSIA*

*Corresponding Author: khuneswari@uthm.edu.my
DOI: https://doi.org/10.30880/ekst.2024.04.02.024

**Article Info**

Received: 27 December 2023
Accepted: 11 January 2024
Available online: 12 December 2024

**Keywords**

Bank Account Fraud, LASSO, Elastic Net, Logistic Regression, K-Fold Cross Validation

**Abstract**

Bank account fraud remains a pervasive threat in the financial industry, resulting in numerous adverse consequences for society. This underscores the critical necessity for the development and implementation of robust and efficient fraud detection mechanisms. The first objective of this study was to compare the performance between LASSO and Elastic Net regression in detecting bank account fraud by using AICc and BIC. The second objective was to choose the best model for bank account fraud detection based on the lowest value of AICc and BIC. Furthermore, this study aimed to identify the factors that affect bank account fraud by using the best model. The dataset used in this study was synthetic data that was published at the Neural Information Processing Systems website. The results indicated that factors affecting bank account fraud included income levels, the age of customers, the duration elapsed since the application was initiated, the initial transferred amount during the application, the credit limit proposed by the applicant, disparities between the origin country of the request and the bank's country, applications submitted through application, the duration of a user's session on the banking website, and the operating system of the device used, encompassing Windows, macOS, X11, and other devices that initiated the request. Besides, logistic Elastic Net regression yielded better results than logistic LASSO regression in bank account fraud detection due to its advanced nature, which combined L1 and L2 regularization techniques. This study suggested that logistic Elastic Net regression is useful for enhancing fraud detection in financial institutions and increasing public awareness by identifying the influential factors that affect fraud.

## 1. Introduction

In recent years, there has been a significant increase in the volume of financial transactions which increases the risk of financial fraud, with reported losses reaching alarming figures. Factors such as technological advancements and the COVID-19 pandemic have provided fraudsters with greater opportunities, resulting in heightened risks for individuals and institutions. Fraudulent transactions which may result in loss of money have become a critical problem in the world. Hence, immediate attention to fraud detection and prevention measures must be implemented. The collaboration between governments, financial institutions, and the public is pivotal in mitigating these risks.

The term “fraud” refers to the illegal acts carried out intentionally to gain personal or organizational benefits by harming other parties [1]. Past research has shown that older people’s excess of trusting nature, social isolation, lack of knowledge in fraud prevention, and other factors might cause the elders to become the target of fraudsters [2]. Fraud may occur in a variety of different forms, with bank account fraud being one of the

most prevalent. Bank account fraud is a specific type of fraud that can occur when a fraudster gains unauthorized access to an individual's bank account which may lead to significant financial losses. Bank account fraud is a pressing issue in society nowadays because it can be carried out by internal or external people associated with banks and might bring serious negative consequences for individuals, financial institutions and families.

Logistic regression is a useful statistical method commonly utilized in fraud detection. Up to now, several studies have used logistic regression in detecting credit card fraud. In fraud detection, logistic regression aims to model outcomes of a class, like passing or failing, positive or negative/neutral. When assessing credit card fraud, it utilizes probability distribution to distinguish between fraudulent and non-fraudulent cases.

LASSO (Least Absolute Shrinkage and Selection Operator) is a regression method that reduces the complexity of a regression model by penalizing the absolute value of a regression coefficient [3]. LASSO is useful in feature selection as the penalty term of LASSO has the effect of forcing some of the coefficient estimates to be exactly equal to zero and choosing the variables that are not zero to stay in the model.

Elastic Net is an advanced method of both ridge regression and LASSO regression which was introduced by Zou and Hastie in the year 2005. The penalty term of the Elastic Net is the combination of ridge regression and LASSO regression. The use of LASSO was demonstrated to be problematic when at least some features are highly correlated and LASSO picks out one feature at random in this circumstance. Elastic Net overcomes the shortcomings and limitations of both ridge regression and LASSO regression. By using this technique, it is possible to create models that consider the multicollinearity of the explanatory variables while simultaneously choosing key characteristics from a significant number of potential predictors.

The purpose of this study is to compare the performance between LASSO and Elastic Net regression in detecting bank account fraud by using AICc and BIC. The best model for bank account fraud detection will be chosen based on the lowest value of AICc and BIC. Lastly, the factors that affect bank account fraud will be identified by using the best model.

## 2. Materials

### 2.1 Data Sources and Data Set

The dataset used in this research is about the Bank Account Fraud suite that has been published at Neural Information Processing Systems. It is a replica of real-world account opening fraud data that was generated using CTGAN (Conditional Tabular Generative Adversarial Network). This data was created with the purpose of the contribution of evaluation employed in ML Research with a large-scale, realistic, and up-to-date suite of tabular datasets. The dataset consists of 1000000 transaction data. There are 32 variables, including income, credit risk, proposed credit limit, and other factors that might affect fraud. It will become a challenge if all of them are included in the research, so only certain significant variables were selected for the development and evaluation of the statistical model intended for fraud detection purposes. These variables were selected based on the existing past research that has examined the significant factors that affect bank account fraud. Certain variables that are not closely related to bank account fraud such as the similarity between email and applicant's name will be eliminated. The dependent variable, $y$ is the occurrence of bank account fraud, and 9 variables will serve as independent variables, $x$ including the annual income of the applicant, the applicant's age, the number of days passed since the application was done, the initial transferred amount for application, 16 applicant's proposed credit limit, origin country of the request is different from the bank's country, online source of application, the length of a user session on the banking website is in minutes and operative system of the device that made the request.

## 3. Methods

### 3.1 Cross Tabulation

Cross tabulation is commonly known as a contingency table and crosstabs. Cross tabulation summarises the relationship between two or more categorical variables in a matrix form which consists of rows and columns that allow the researchers to observe the patterns or dependencies between variables simultaneously. For the 2x2 contingency table, each row indicates the categories of one variable and the columns represent the categories of another variable. In this case, the collected data were analysed using cross-tabulation analysis to identify categorical variables that show a significant association with fraud occurrences. Cross tabulation is employed to examine the patterns and relationships between different variables related to fraud cases.

#### 3.1.1 Chi-Square Test

The Chi-Square test and cross-tabulation are frequently used in conjunction as a thorough method for assessing the relationship between categorical variables in a contingency table. The Chi-Square test enables a comparison

between observed and expected frequencies [4]. According to [5], the formula of the Chi-Square test is shown in equation (1).

$$\chi^2 = \frac{(\sum (O_i - E_i)^2)}{E_i} \quad (1)$$

where $X^2$ indicates the test statistics, $O_i$ represents the observed value for $i^{th}$ observations, and $E_i$ is the expected value for $i^{th}$ observations. The null hypothesis of the Chi-Square test stated that there is no association between the categorical variables, while the alternative hypothesis stated that there is an association between the categorical variables. The null hypothesis will be rejected if the *p*-value is less than the significance level.

### 3.1.2 Data Pre-processing

Data preprocessing is an essential initial step to ensure data quality before conducting analyses. Raw data often contains noise, corruption, missing values, and inconsistencies that can significantly impact result accuracy and reliability [6]. Common methods in data preprocessing, like outlier elimination, data decomposition, and denoising, aim to address these issues. Handling missing values typically involves deletion or imputation techniques. For instance, in our study, missing values in variables such as user session lengths on a banking website were imputed by computing the mean or median of the non-missing values within the same variables.

### 3.1.3 K-Fold Cross-Validation

K-fold cross-validation is an essential tool in machine learning that can help in performance evaluation and parameter selection [7]. K-fold cross-validation performed in this study serves two purposes. The first purpose is to split the data into training and testing data based on the lowest mean squared error. While the second purpose is to choose the optimal tuning parameters. Regularization regression methods such as ridge regression, LASSO regression, and Elastic Net regression usually rely on tuning parameters [8]. A tuning parameter denoted as $\lambda$ (lambda) controls the degree or strength of the penalty term applied to coefficient estimates and the type of penalization. There are a few approaches to selecting the best tuning parameter. One of the most popular methods is using K-fold cross-validation to calculate the cross-validation error and choose the optimal lambda based on the regularization path with the lowest cross-validation error.

However, according to the study by [3], binomial deviance is used as an evaluation metric instead of mean squared error for logistic regression. Hence, since this research focuses on logistic regression, the optimal value of tuning parameters (lambda and alpha) will be chosen based on the minimum value of binomial deviance, which is more appropriate for this type of regression. This process is repeated k times and the error of each lambda value will be recorded. In this research, the value of $k$ = 10 will be employed for K-fold cross-validation.

### 3.1.4 Logistic Regression

According to [9], logistic regression is applied to describe the relationship between binary outcomes and one or more independent variables. The objective of logistic regression is to estimate the probability of a binary outcome for a dependent variable based on the independent variables. In this case, the outcome variable is the occurrence of fraud and has two categories which are “yes” and “no”. The occurrence of an event is encoded as one for “yes” and zero as “no”.

### 3.1.5 Least Absolute Shrinkage and Selection Operator (LASSO)

The objective of LASSO regression is to minimize the cost function by identifying the values of coefficients ($\beta$) that balance the trade-off between mean squared error (MSE) and the L1 penalty. The LASSO approach is preferred by many researchers because this method allows researchers to select important variables by choosing the non-zero values of the variables through L1 regularization and excluding those unimportant variables [10]. The formula of LASSO is shown in equation (2) [11].

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \sum_{j=0}^{p} B_j \times x_{ij})^2 + \lambda \sum_{j=0}^{p} |B_j| \quad (2)$$

where $n$ represents a number of observations, $p$ indicates the number of predictor variables, $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the sum of squared error (SSE), $\sum_{i=1}^{n}(y_i - \sum_{j=0}^{p} B_j \times x_{ij})^2$ is the sum of squared residuals (SSR), and $\lambda \sum_{j=0}^{p} |B_j|$ represents the sum of absolute values of the magnitude of the coefficients (L1 penalty). However, it is crucial to select

optimal lambda for the regularization method because lambda controls the amount of regularization in the model which may affect the bias-variance-trade-off. As lambda λ increases, the L1 penalty increases, which can cause the coefficients to shrink toward zero and eliminate those unimportant variables. If the lambda chosen is too small may lead to overfitting, where the solution is overly optimized, resulting in high variance [12]. If the chosen lambda is too large, the solution will be regular but have an unacceptably large objective value associated with the objective function. Hence, the regularization path is introduced to address this issue. An optimal lambda can be chosen based on the properties and features such as elbow point visualized from the regularization path.

### 3.1.6 Elastic Net

Elastic Net regression is an advanced technique to overcome the shortcomings of LASSO regression. Elastic Net regression is suitable when the number of predictor variables (*p*) is larger than the number of response variables (*n*), and when the data are multicollinear [13]. The objective of Elastic Net regression is like LASSO regression, which is to minimize the cost function by identifying the values of coefficients ($\beta$) that balance the trade-off between mean squared error (MSE) and the L1 penalty and L2 penalty. According to a study by [11], the formula of Elastic Net regression is shown in equation (3).

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(y_i-\sum_{j=0}^{p}B_j\times x_{ij})^2+\lambda\sum_{j=0}^{p}\left|B_j\right|+\lambda\sum_{j=0}^{p}B_j{}^2 \tag{3}$$

where *n* represents number of observations, *p* indicates the number of predictor variables, $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ is the sum of squared error, $\sum_{i=1}^{n}(y_i-\sum_{j=0}^{p}B_j\times x_{ij})^2$ is the sum of squared residuals (SSR), $\lambda\sum_{j=0}^{p}\left|B_j\right|$ represents the sum of absolute values of the magnitude of the coefficients (L1 penalty), and $\lambda\sum_{j=0}^{p}B_j{}^2$ indicates the sum of the squared values of the magnitude of the coefficients (L2 penalty). As λ increases, the L1 and L2 penalties increase, which can reduce the impact of unimportant variables in the model. This is how logistic Elastic Net regression helps to reduce the complexity and overfitting of the model. The selection of optimal lambda for Elastic Net regression is the same as LASSO regression. The regularization path will be employed for the optimal lambda for Elastic Net regression.

### 3.1.7 Model Selection Criteria

The two criteria, the Akaike Information Criterion corrected (AICc) and Bayesian Information Criterion (BIC) will serve as the standardized evaluation criteria for model selection in this research. The modified version of AIC which is AICc was used in this research instead of AIC. The reason is that AICc considers the sample size to reduce model complexity with small data sets by increasing the relative penalty. In contrast to AIC, which only considers the number of parameters, both AICc and BIC consider the interplay between the number of parameters and the sample size (*n*). Equation (4) represents the AIC formula, while equation 5 represents the AICc formula [14]. The model with the lowest AICc and BIC will be considered the best-fitting model [15].

$$AIC=-2\log\ell(\hat{\theta})+2k \tag{4}$$

$$AICc=AIC+\frac{2k(k+1)}{n-k-1} \tag{5}$$

Bayesian Information Criterion (BIC) was developed in 1978. Both AIC and BIC are useful for model selection and balance the complexity and the accuracy of the model by adding a penalty term to the likelihood function [16]. However, they differ in the penalty terms they use. In the Bayesian Information Criterion (BIC), the impact of the number of model parameters on the penalty is determined by considering the sample size "*n*", which is shown in equation (6). On the other hand, in the Akaike Information Criterion (AIC), the sample size does not play a role in determining the penalty, it considers the number of parameters only. Consequently, when the sample size is small, AIC will likely select models that include a larger number of parameters. The formula of BIC is shown in equation (6) [14].

$$BIC=-2\log\ell(\hat{\theta})+k\ln(n) \tag{6}$$

Enhanced Knowledge in Sciences and Technology Vol. 4 No. 2 (2024) p. 201-211

where -2log$\ell(\theta)$ is the log-likelihood, $n$ represents the sample size of the data, and $k$ is the number of parameters.

### 3.1.8 Goodness of Fit Test

The goodness of fit test is applied to determine how well the model fits the data by checking the assumptions of the model. In this research, both deviance residual plots and the statistical method, the Chi-Square test will be utilized to assess the goodness of fit test.

#### 3.1.8.1 Deviance Residual Plot

Due to LASSO and Elastic Net regression being generalized linear models (GLM), the deviance residual plot is used to evaluate the model suitability of LASSO regression and Elastic Net regression. The deviance residuals plot is plotted to observe the patterns. The residual analysis helps in evaluating whether the model is a good fit or not through visualization. A residual plot with a random pattern, constant variance (homoscedasticity) and normally distributed residuals with zero means indicate a good fit.

#### 3.1.8.2 Chi-Square Test

The deviance residual plot serves as a visualization tool to assess the model's fit to the data. However, to validate this fit statistically, methods like the Chi-Square test are essential. Therefore, it is crucial to construct a null model containing only the intercept and compare it with the best logistic model. The deviance residuals are computed based on the deviance, where the test statistics are calculated as the difference between the log-likelihood fitted model and the maximum likelihood of the chosen saturated model [17]. The deviance residual can be calculated by using the equation (7) [18].

$$r_D = sign(y_i - \hat{\mu}_i)\sqrt{D} \tag{7}$$

where $rD$ represents the deviance residual, $y_i$ is the observed value for the $i^{th}$ observation, $\hat{u}_i$ is the predicted mean value for the $i^{th}$ observation and $D$ is the deviance function. The deviance function can be computed based on the formula shown in equation (8) [18].

$$D = 2\sum[y_i \log(\frac{y_i}{\hat{\mu}_i}) - (y_i - \hat{\mu}_i) \tag{8}$$

By computing the deviance difference of both models, the resulting $p$-value determines significance. The null hypothesis implied that the model perfectly fits the data, while the alternative hypothesis indicated that the model does not adequately fit the data. Consequently, if the computed $p$-value falls below the significance level of 0.05, the null hypothesis will be rejected.

## 4. Results and Discussion

### 4.1 Data Pre-processing

An analysis must always be preceded by data pre-processing. The initial phase involved examining the missing data as missing data can significantly impact the quality of any analysis. In this dataset, two variables exhibited missing values, namely "intended_balcon_amount" and "session_length_in_minutes." The variable "intended_balcon_amount" had 742522 missing values, whereas the variable "session_length_in_minutes" had 2015 missing values. Due to the high occurrence of missing values in both quantitative variables, mean imputation was employed to replace all missing values. In addition, this dataset included three categorical variables, which are “foreign_request”, “source” and “device_os”. To prepare these variables for subsequent analysis, all the categorical variable columns were substituted using label encoding with numerical values, and the encoded categorical variable columns were then converted to factors.

### 4.2 Cross Tabulation and Chi-Square Test

After completing all the steps in data pre-processing, cross-tabulation, and Chi-Square tests were employed to assess associations between categorical data. The results were presented in tabular form. Table 1 illustrates the $p$-value obtained by the Chi-Square test for the 'foreign_request', ‘source’ and ‘device_os’ variables. The associated $p$-value for all the variables was 0.000, which was less than the significance level of 0.05.

Consequently, the null hypothesis, suggesting there was no association between the variables "fraud_bool" and "foreign_request," was rejected. Besides that, there was a significant relationship between the "fraud_bool" variable and the "source" variable. Lastly, the results suggested there was an association between "fraud_bool" and "device_os".

**Table 1** The p-value obtained by Chi-square test

| Variables | p-value |
|---|---|
| Foreign request | 0.000 |
| Source | 0.000 |
| Device_os | 0.000 |

## 4.3 K-Fold Cross-Validation

K-fold cross-validation was performed to serve two distinct purposes. Firstly, the dataset was split into training data and testing data, with the value of *k* set at 10, meaning the data was divided into ten equal folds. The model was trained on the remaining *k*-1 folds. This process was repeated *k* times, which was ten times in this case. Mean-squared error (MSE) served as the standard metric for evaluating the performance of the testing data in each fold. The fold with the lowest mean-squared error was chosen as the best model because its prediction was closest to the actual value. Fold three was identified as the best fold with the lowest mean squared error of 0.01103. The second purpose of K-fold cross-validation was to determine the optimal tuning parameters. Both logistic LASSO regression and logistic Elastic Net regression involved determining optimal lambda. K-fold cross-validation was used to find the optimal alpha for logistic Elastic Net regression. However, K-fold cross-validation chose the training data with the lowest mean squared error. Since this was a logistic regression model, optimal lambda and alpha were chosen based on minimum binomial deviance.

## 4.4 Regression Model

Logistic LASSO regression was implemented for feature selection. To perform logistic LASSO regularization, the regularization parameter, alpha was set to one, indicating a strong L1 regularization effect. Some of the coefficients were shrunk to be exactly zero during the regularization process. This step aimed to select the most informative features.

As discussed earlier, K-fold cross-validation was utilized to choose the optimal lambda. The "glmnet" package was employed for building linear and logistic regression models. The value of 0.00005897 represented the optimal lambda when all 12 features were selected. Once the optimal lambda was identified, the final model was trained using this chosen lambda. The logistic LASSO regularization path, as visualized in Fig.1 (a), illustrated the effect of regularization, with lambda serving as the tuning parameter. The optimal lambda obtained through K-fold cross-validation corresponded to the specific point on the regularization path. Furthermore, Fig.1 (a) provided a clear illustration of the lambda effect on the model's coefficients. Higher lambda values indicated stronger regularization, resulting in a stronger tendency of reduction in the magnitude of coefficients, with some coefficients shrinking towards zero. This reduction simplified the model and decreased its complexity.

Logistic Elastic Net regression combined the principles of both Ridge regression (L2 regularization) and Lasso regression (L1 regularization). Since the penalty term, alpha hyperparameter of logistic Elastic Net regression was a mixture of Ridge regression and LASSO regression, the optimal alpha had to lie between the range of optimal alpha for both logistic Ridge regression and logistic LASSO regression. Therefore, the optimal alpha value was set in the range between 0.1 to 0.9. To determine the optimal alpha and lambda values for the model, K-fold cross-validation was employed using the "cv.glmnet" function in R. The selection of the optimal alpha and lambda was driven by minimizing the binomial deviance. The best alpha value, which balanced between logistic Ridge regression and logistic LASSO regression, was found to be 0.2. The optimal lambda, which controlled the strength of regularization, achieved a balance between sparsity and regularization strength at 0.0001537.

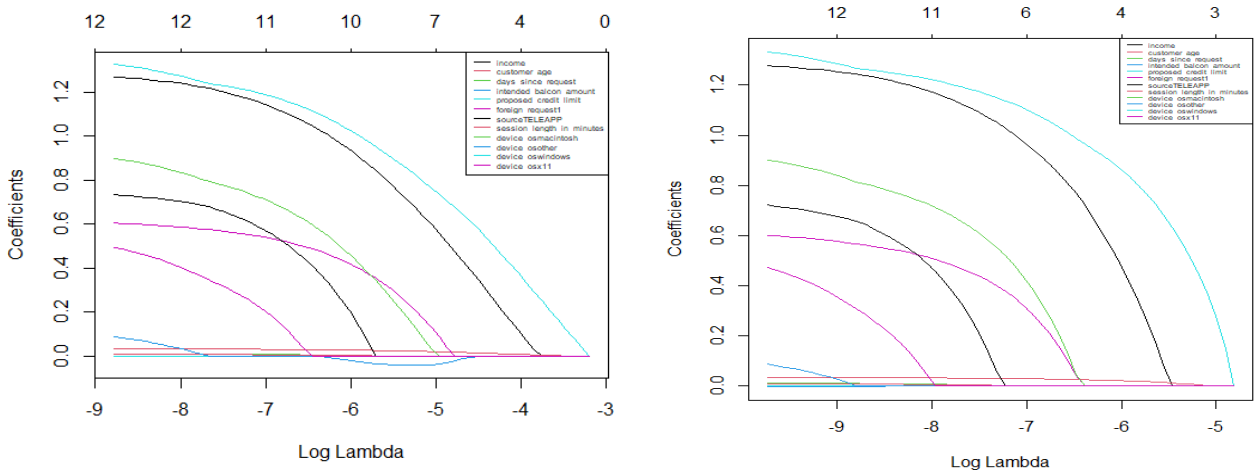
Logistic Elastic Net regression worked similarly to logistic LASSO regression after the optimal alpha and lambda values were found. The final model was trained using the optimal lambda and alpha values that were identified. The logistic Elastic Net regularization path is shown in Fig.1 (b). This graphic provided insightful information on how different regularization strengths affect these coefficients' modifications. Stronger regularization led to more coefficients being penalized and shrinking toward zero, as visualized in Fig.1 (b), which enhanced the model's simplicity and sparsity.

Enhanced Knowledge in Sciences and Technology Vol. 4 No. 2 (2024) p. 201-211

**Table 2** Coefficients of Predictor Variables in Logistic LASSO regression

| Predictor Variable | LASSO | Elastic Net |
|---|---|---|
| Intercept | -7.6712 | -7.6706 |
| Income_$X_1$ | 1.2772 | 1.2764 |
| Customer_age_ $X_2$ | 0.0334 | 0.0333 |
| Days_since_request_$X_3$ | 0.0111 | 0.0112 |
| Intended_balcon_amount_$X_4$ | -0.0025 | -0.0026 |
| Proposed_credit_limit_$X_5$ | 0.0007 | 0.0007 |
| Foreign_request1_$X_6$ | 0.6003 | 0.6027 |
| sourceTELEAPP_$X_7$ | 0.7206 | 0.7256 |
| Session_lengths-in_minutes_$X_8$ | 0.0079 | 0.0079 |
| Device_osmacintosh_$X_9$ | 0.9036 | 0.9062 |
| Device_osother_$X_9$ | 0.0873 | 0.0910 |
| Device_oswindows_$X_9$ | 1.3335 | 1.3342 |
| Device_osx11_$X_9$ | 0.4723 | 0.4833 |

The resulting coefficients for both the logistic LASSO regression and logistic Elastic Net regression are presented in Table 2. The logistic Elastic Net regression model retained the same variables as those in the logistic Lasso regression. Additionally, the coefficients of logistic Elastic Net regression did not exhibit significant variation compared to logistic LASSO regression. The exclusion of the three variables, namely "foreign_request0", "sourceINTERNET", and "device_oslinux" led to an understanding of the factors that affected the possibility of fraud occurrence. It can be concluded that the absence of foreign requests, transactions conducted via the internet, and the use of Linux devices to perform transactions had negligible influence on predicting the occurrence of fraud. This result underscored the consistency in variable selection between the two regression models.




**(a)** **(b)**

**Fig. 1** (a) Logistic LASSO regularization path; (b) Logistic Elastic Net regularization path

The model equation for both logistic LASSO regression and logistic Elastic Net regression were outlined in equation (9) and equation (10). Where *P* represents the probability of fraud occurrence.

$$P(FraudOccurrence) = \frac{exp\begin{pmatrix} -7.6712 + 1.2772 * Income + 0.0334 * customer\_age + 0.0111 * days\_since\_request - 0.0025 * intended\_balcon\_amount \\ +0.0007 * proposed\_credit\_limit + 0.6003 * foreign\_request1 + 0.7206 * sourceTELEAPP + 0.0079 * session\_lengths\_in\_minutes \\ +0.9036 * device\_osmacintosh + 0.0873 * device\_osother + 1.3335 * device\_oswindows + 0.4723 * device\_osx11 \end{pmatrix}}{1 + exp\begin{pmatrix} -7.6712 + 1.2772 * Income + 0.0334 * customer\_age + 0.0111 * days\_since\_request - 0.0025 * intended\_balcon\_amount \\ +0.0007 * proposed\_credit\_limit + 0.6003 * foreign\_request1 + 0.7206 * sourceTELEAPP + 0.0079 * session\_lengths\_in\_minutes \\ +0.9036 * device\_osmacintosh + 0.0873 * device\_osother + 1.3335 * device\_oswindows + 0.4723 * device\_osx11 \end{pmatrix}} \quad (9)$$

$$P(FraudOccurrence) = \frac{exp\begin{pmatrix} -7.6706 + 1.2764 * Income + 0.0333 * customer\_age + 0.0112 * days\_since\_request - 0.0026 * intended\_balcon\_amount \\ +0.0007 * proposed\_credit\_limit + 0.6027 * foreign\_request1 + 0.7256 * sourceTELEAPP + 0.0079 * session\_lengths\_in\_minutes \\ +0.9062 * device\_osmacintosh + 0.0910 * device\_osother + 1.3342 * device\_oswindows + 0.4833 * device\_osx11 \end{pmatrix}}{1 + exp\begin{pmatrix} -7.6706 + 1.2764 * Income + 0.0333 * customer\_age + 0.0112 * days\_since\_request - 0.0026 * intended\_balcon\_amount \\ +0.0007 * proposed\_credit\_limit + 0.6027 * foreign\_request1 + 0.7256 * sourceTELEAPP + 0.0079 * session\_lengths\_in\_minutes \\ +0.9062 * device\_osmacintosh + 0.0910 * device\_osother + 1.3342 * device\_oswindows + 0.4833 * device\_osx11 \end{pmatrix}} \quad (10)$$

## 4.5 Comparison of Logistic LASSO Regression and Logistic Elastic Net Regression

**Table 3** Comparison of Regression Models Based on AICc and BIC scores

| Regression Models | AICc | BIC |
|---|---|---|
| Logistic LASSO | −11082.99 | −10942.47 |
| Logistic Elastic Net | −11083.57 | −10943.05 |

The AICc and BIC values for both regression methods are presented in Table 3. According to Table 3, it was noted that AICc and BIC values for both logistic LASSO regression and logistic Elastic Net regression were quite similar, with minor differences observed in the decimal places. However, these minor differences still provided valuable insights. Many journals stated that Elastic Net regression performed better than LASSO regression due to Elastic Net regression's ability to strike a balance between feature selection and coefficient shrinkage. Despite a little discrepancy in the evaluation parameters, Elastic Net regression still outperformed LASSO regression [19]. Smaller AICc and BIC values indicated better performance of the model in the context of bank account fraud detection. A conclusion was drawn that logistic Elastic Net regression outperformed logistic LASSO regression since it had smaller AICc and BIC values. Hence, logistic Elastic Net regression was considered the best model for bank account fraud detection.

## 4.6 Goodness of Fit Test

### Deviance Residual Plot

The deviance residual plot was useful in assessing the performance of the model by visualization. The trends or patterns provided insights to see whether the model was a good fit. In Fig.2, the deviance plot of the logistic Elastic Net regression model is presented, showcasing deviance residuals plotted against predicted probabilities (fitted values). As visualized in Fig.2, there were two clear parallel lines in the plot. Particularly, the upper line (coded as one) represented cases where fraud events were observed, while the bottom line (coded as zero) showed the events where fraud cases were absent. The presence of these parallel lines was an important finding as they demonstrated that the model operated steadily and consistently throughout a range of predicted probabilities. Hence, the deviance residual plot suggests that the model was a good fit for the data. However, it is important to conduct a *p*-value test to fully assess the performance of the model on the data.

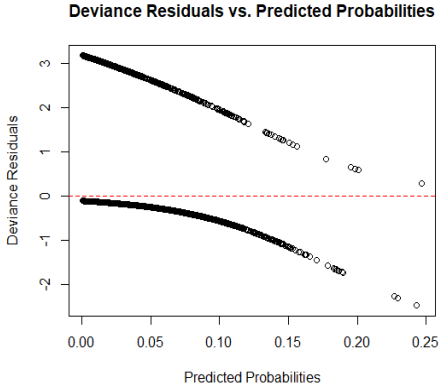



**Fig.** 2 Deviance Plot of logistic Elastic Net regression

### 4.6.1 Chi-Square Test

**Table 4** Comparison of Model Residuals

| Model | Value of Residuals | Degree of Freedoms |
|---|---|---|

| | | |
|---|---|---|
| Null Model | 12136.53 | 1 |
| Logistic Elastic Net | 11002.22 | 13 |

Based on the results obtained from the comparison between both regressions, it was clearly shown that logistic Elastic Net regression performed better when compared to logistic LASSO regression. Hence, the testing data of logistic Elastic Net regression was used to conduct the deviance goodness of fit test. The evaluation method commenced by creating a null model that contained only an intercept term. The main objective of this null model was to serve as a reference point based on its null deviance. Subsequently, the reference value was compared to the deviance of the actual model, which was the logistic Elastic Net regression. The Chi-Squared statistics value was determined by computing the difference between the null deviance and the residual deviance, considering the appropriate degrees of freedom. The Chi-Squared statistics value was found to be 1134.315. A larger Chi-Squared statistics value indicated a better fit of the data. The *p*-value was calculated on the Chi-Squared statistics equal to 1, which was larger than the significance value, 0.05. As a result, the null hypothesis, which suggested that the model perfectly fits the data was accepted.

## 5. Conclusion

The research successfully addressed the three specified objectives by employing both logistic LASSO and logistic Elastic Net regression methods. In this research, the comparison of logistic LASSO regression and logistic Elastic Net regression in detecting bank account fraud using AICc and BIC provided meaningful insights into their respective performances. AICc and BIC served as the standard evaluation metrics to evaluate the two regression models, which were logistic LASSO regression and logistic Elastic Net regression. The results showed that Elastic Net regression outperformed logistic LASSO regression because logistic Elastic Net regression had the lowest value for both AICc and BIC. Despite the values for both regressions being close to each other, the goodness of fit test indicated that logistic Elastic Net regression was a suitable model fit for the data. This superiority is attributed to Elastic Net's effectiveness in handling high-dimensional and correlated data compared to logistic LASSO regression [20]. Moreover, the identified factors influencing bank account fraud included income levels, the age of customers, the duration elapsed since the application was initiated, the initial transferred amount during application, the credit limit proposed by the applicant, disparities between the origin country of the request and the bank's country, request submitted through application, the duration of a user's session on the banking website, and the operating system of the device used, encompassing Windows, macOS, X11, and other devices that initiated the request. In conclusion, the study successfully achieved all three objectives, providing valuable insights for government and banking institutions. This research made a substantial contribution to a better understanding and offered financial institutions the best model, logistic Elastic Net regression, to enhance their systems in fraud detection and prevention. Furthermore, this study successfully contributed to enhancing public understanding of bank account fraud by identifying significant contributing factors to fraud.

The study encountered limitations, primarily in data availability for logistic LASSO and Elastic Net regression in addressing fraud, emphasizing the widespread concern surrounding this issue. However, the availability of fraud data was limited, primarily due to the presence of confidential information. Besides that, it was challenging to doubt or raise concerns about data related to fraud when it was supposedly gathered in a manner that kept the details confidential and anonymous. Hence, it was recommended to advocate for collaborative efforts between government and research institutions to facilitate the sharing of fraud data.

## Acknowledgement

First, I would like to express my sincere gratitude and appreciation to my supervisor, Dr. Khuneswari A/P Gopal Pillay for allowing me to conduct this research and providing encouragement, valuable insights and guidance to me throughout the research. Besides that, I would like to thank my university, Universiti Tun Hussein Onn Malaysia for allowing me to participate in this course and carry out this research. In addition, I am deeply grateful for the support and help from my beloved friends. This research wouldn't have able completed without their support. Their insightful suggestion and selfless sharing of knowledge have played a crucial role in the completion of this research. Lastly, I would like to convey thanks to every individual including all the lecturers who have helped me along the way. Thank you for standing by my side in my journey of development.

## Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of the paper.

## Author Contribution

*The authors confirm their contribution to the paper as follows:* ***study conception and design:*** *Tan Jing Thong, Khuneswari Gopal Pillay;* ***data collection:*** *Tan Jing Thong;* ***analysis and interpretation of results:*** *Tan Jing*

*Thong, Khuneswari Gopal Pillay;* ***draft manuscript preparation:*** *Tan Jing Thong, Khuneswari Gopal Pillay. All authors reviewed the results and approved the final version of the manuscript.*

## References


[1] Ballo, B. E. R., Dethan, M. A., Angi, Y. F., & Rafael, S. J. M. (2023). Analysis of Tone at the Top Principle's Implementation in Fraud Prevention on Public University in Kupang City. Asia Pacific Fraud Journal, 8(1), 1. https://doi.org/10.21532/apfjournal.v8i1.251

[2] Shao, J., Zhang, Q., Ren, Y., Li, X., & Lin, T. (2019). Why are older adult's victims of fraud? Current knowledge and prospects regarding older adults' vulnerability to fraud. Journal of Elder Abuse & Neglect, 31(3), 225–243. https://doi.org/10.1080/08946566.2019.1625842

[3] Kawase, R., Diana, F., Czeladka, M., Schüler, M., & Faust, M. (2019). Internet Fraud: The Case of Account Takeover in Online Marketplace. https://doi.org/10.1145/3342220.3343651

[4] Bortolini, R., & Forcada, N. (2020). Analysis of building maintenance requests using a text mining approach: building services evaluation. Building Research and Information, 48(2), 207–217. https://doi.org/10.1080/09613218.2019.1609291

[5] Maharana, K., Mondal, S., & Nemade, B. P. (2022). A review: Data pre-processing and data augmentation techniques. Global Transitions Proceedings, 3(1), 91–99. https://doi.org/10.1016/j.gltp.2022.04.020

[6] McHugh, M. M. (2013). The Chi-square test of independence. *Biochemia Medica*, 143–149. https://doi.org/10.11613/BM.2013.018

[7] Li, J., Gao, F., Lin, S., Guo, M., Li, Y., Liu, H., Qin, S., & Wen, Q. G. (2022). Quantum k□fold cross-validation for nearest neighbor classification algorithm. Physica D: Nonlinear Phenomena, 611, 128435. https://doi.org/10.1016/j.physa.2022.128435

[8] Ahrens, A., Hansen, C., & Schaffer, M. E. (2020). lassopack: Model selection and prediction with regularized regression in Stata. Stata Journal, 20(1), 176–235. https://doi.org/10.1177/1536867x20909697

[9] Schober, P., & Vetter, T. (2021). Logistic Regression in Medical Research. Anesthesia & Analgesia, 132(2), 365–366. https://doi.org/10.1213/ane.0000000000005247

[10] Lee, J. H., Shi, Z., & Gao, Z. (2021). On LASSO for predictive regression. Journal of Econometrics, 229(2), 322–349. https://doi.org/10.1016/j.jeconom.2021.02.002

[11] Nieto, P. G., García–Gonzalo, E., & Paredes–Sánchez, J. P. (2021). Prediction of the critical temperature of a superconductor by using the WOA/MARS, Ridge, Lasso and Elastic-net machine learning techniques. Neural Computing and Applications, 33(24), 17131–17145. https://doi.org/10.1007/s00521-021-06304-z

[12] Gebken, B., Bieker, K., & Peitz, S. (2022). On the structure of regularization paths for piecewise differentiable regularization terms. Journal of Global Optimization, 85(3), 709–741. https://doi.org/10.48550/arxiv.2111.06775

[13] Wang, W., Liang, J., Liu, R., Song, Y., & Zhang, M. (2022). A robust variable selection method for sparse online regression via the elastic net penalty. Mathematics, 10(16), 2985. https://doi.org/10.3390/math10162985

[14] Hendrawati, T., Wigena, A. H., Sumertajaya, I. M., & Sartono, B. (2021). Clustering of Commodity Inflation Pattern based on Estimated ARIMA Model. Journal of Physics: Conference Series, 1863(1), 012058. https://doi.org/10.1088/1742-6596/1863/1/012058

[15] Nye, C. D., Joo, S., Zhang, B., & Stark, S. (2019). Advancing and evaluating IRT model data fit indices in organizational research. Organizational Research Methods, 23(3), 457–486. https://doi.org/10.1177/1094428119833158

[16] Zhang, Y., & Meng, G. (2023). Simulation of an Adaptive Model Based on AIC and BIC ARIMA Predictions. Journal of Physics, 2449(1), 012027. https://doi.org/10.1088/1742-6596/2449/1/012027

[17] Chandrakantha, L. (2019). Risk Prediction Model for Dengue Transmission Based on Climate Data: Logistic Regression Approach. Stats, 2(2), 272–283. https://doi.org/10.3390/stats2020021

[18] Khan, A. S., Ullah, M. W., Amin, M., Muse, A. H., Aldallal, R., & Mohamed, M. A. (2022). Empirical Examination of the Poisson Regression Residuals for the Evaluation of Influential Points. Mathematical Problems in Engineering, 2022, 1–9. https://doi.org/10.1155/2022/6995911

Enhanced Knowledge in Sciences and Technology Vol. 4 No. 2 (2024) p. 201-211

[19] García-Carretero, R., Vigil-Medina, L., Barquero-Pérez, Ó., Jiménez, I., Soguero-Ruíz, C., Goya–Esteban, R., & Ramos-López, J. (2020). Logistic LASSO and Elastic Net to characterize vitamin D deficiency in a hypertensive obese population. Metabolic Syndrome and Related Disorders, 18(2), 79–85. https://doi.org/10.1089/met.2019.0104
[20] Hu, Y. (2022). Applications of Elastic Net technology in survival analysis of high‐dimensional data. 2022 4th International Conference on Big Data Engineering. https://doi.org/10.1145/3538950.3538955