Machine Learning Classifiers for Detecting Credit Card Fraudulent Transactions

Bharti Chugh and Nitin Malik

Abstract Credit card usage has increased significantly as a result of the fast development of e-commerce and the Internet. As a consequence of enhanced credit card usage, credit card theft has risen substantially in recent years. Fraud in the financial sector is expected to have far-reaching effects in the near future. As a response, numerous scholars are concerned with financial fraud detection and prevention. In order to prevent bothering innocent consumers while detecting fraud, accuracy has become critical. We used hyperparameter optimization to see if created models utilizing different machine learning approaches are significantly the same or different, and if resampling strategies improve the suggested models' performance. The hyperparameter is optimized using GridSearchCV techniques. To test the hypotheses of data that has been divided into training and test data, the GridSearchCV and random search methods are used. The maximum accuracy 72.1% was achieved by decision tree classifier on the imbalanced German credit card dataset. The maximum accuracy of 98.6% is achieved by LDA on imbalanced European credit card dataset. Additionally, logistic regression and naïve Bayes were also tested and SMOTE was applied.

Keywords Decision tree · LDA · Gaussian Naïve Bayes · Logistic regression · Bernoulli Naïve Bayes · Credit card · GridSearchCV

1 Introduction

Major financial institutions are making their services available to the general public through online banking, mobile banking, credit and debit cards. Using services like credit cards, which have proven to be extremely good way for online purchases, makes everyday life easier. In the banking industry, credit card and online net banking fraud is a global issue. The credit card or any other card for the matter has data stored in a machine-readable format on a black stripe on the back. It includes details like the

B. Chugh (✉) · N. Malik
The NorthCap University, Gurugram, India
e-mail: bharti.kathpalia@gmail.com

A. Joshi et al. (eds.), *Information and Communication Technology for Competitive Strategies (ICTCS 2021)*, Lecture Notes in Networks and Systems 400,
https://doi.org/10.1007/978-981-19-0095-2_23

cardholder's name, card number, expiration date, CVV code, card type, and other information that might be used to conduct credit card fraud. In the discipline of fraud detection with classifiers, financial fraud is a major issue. The assumption of balanced distribution in the dataset [1, 2] is a challenge for virtually all classifier learning methods. Machine learning techniques are used to anticipate various bank transactions. This article investigates the effectiveness of machine learning classifiers: logistic regression (LR), linear discriminant analysis (LDA), naïve Bayes (NB), and decision tree (DT). The algorithm's performance is evaluated using the recall score, accuracy, f1-score, and precision.

1.1 Literature Review

The default rate on credit loans across all commercial banks has been at an all time since last six years, according to Federal Reserve economic statistics, and it is expected to continue to rise into 2021. Duman et al. [3] sought to demonstrate the benefits of using data mining methods such as DT and support vector machines (SVM) to solve the credit card fraud recognition problem and reduce the banks' risk. The findings revealed that classifiers and added DT classifiers outperformed SVM approaches in tackling the problematic at hand. Wang et al. [4] presented a strategy for detecting credit card fraud based on local isolation coefficient to mining distance-based outliers on conventional algorithms. Bhattacharya et al. [5] detailed a comparative research on data mining methodologies but with the limitation of non-availability of exact time stamp data beyond the date of credit card transactions. APATE [6] is a new technique for detecting fraudulent credit card transactions in highly nonlinear models. A behavior-based credit card fraud detection model was proposed by Zhang et al. [7]. Chuang et al. [8] created a data mining-based model. Web services were utilized to communicate a fraud design, data communication between banks used to detect fraud. To identify credit card thefts, Yu et al. [9] suggested an outlier mining technique. Dembrani et al. [10, 11] proposed a comparative analysis of various adaptive filter structures that can be executed for credit card fraud recognition. A fusion method was presented [12, 13]. The four components were a rule-based filter, a Dempster-Shafer Adder, a transaction history database, and a Bayesian learner. Srivastava et al. [14–16] developed a hidden Markov model for detecting credit card fraud. They developed a unique credit card fraud detection system that uses best matching algorithms to detect 4 distinct patterns of fraud cases and addresses the associated difficulties reported by previous credit card fraud detection studies [17–19].

1.2 Organization of the Paper

The machine learning techniques applied to the proposed model are explained in Sect. 2. Section 3 depicts the suggested model's block diagram, flowchart, and entire implementation. Section 4 illustrates the comparative study with existing machine learning approaches. Section 5 discusses the conclusion and future scope.

2 Machine Learning Algorithms

2.1 Logistic Regression

The logistic regression model calculates a weighted sum of input characteristics and bias. Logistic regression is named for the function used at the core of the method, the logistic function. Any integer with a real value can be translated to a value between 0 and 1.The output value being modeled is a binary value (0 or 1) rather than a numeric number, which is a major distinction from linear regression. The logistic regression equation is shown below:

$$y = \frac{e^{b0+b1*x}}{1 + e^{b0+b1*x}} \tag{1}$$

where y is the expected output, $b0$ represents the bias or intercept term, and $b1$ represents the coefficient for a single input value (x). Each column in your input data has a b coefficient (a constant real number) that must be determined using your training data.

2.2 Decision Tree

Decision tree (DT) is a non-parametric supervised learning approach used for classification and regression. The objective is to learn basic decision rules from data characteristics to construct a model that predicts the class of a target variable. Instances are classified using decision trees by sorting them along the tree from the root to a leaf node, which yields the classification. Starting at the root node of the tree, an instance is categorized by testing the attribute given by its node, then proceeding along the tree branch according to the attribute's value. The sub-tree rooted at the new node is then processed in the same way.

2.3 Naïve Bayes

The naïve Bayes algorithm utilizes the Bayes theorem to classify the data. The naïve Bayes method essentially tells us the likelihood of a record belonging to a definite class constructed on the standards of its characteristics. Gaussian NB is a form of naïve Bayes that handles continuous data and follows the Gaussian normal distribution.

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(x_i - \mu_y\right)^2}{2\sigma_y^2}\right) \tag{2}$$

The parameters σ_y and μ_y are estimated using maximum likelihood.

The Bernoulli NB decision rule is based on the Bernoulli naïve Bayes decision rule:

$$P(x_i|y) = P(i|y)x_i + (1 - P(i|y))(1 - x_i) \tag{3}$$

2.4 Linear Discriminant Analysis (LDA)

The LDA model assumes that the data is normally distributed and estimates the mean and variance for each class. It is common to assume about this in the multivariate (single input parameter) case with two classes. Overall mean (μ) number of each input (x) for each class (k) may be found by dividing the sum of values by the total number of values.

$$k = \frac{1}{nk}(\text{sum}(x)) \tag{4}$$

The numeral of events with class k is nk, and the mean value of x for class k is μk. The variance (σ^2) is calculated to average squared modification of all value from the mean:

$$\sigma^2 = \frac{1}{(n-k)}\text{sum}\left((x-)^2\right) \tag{5}$$

3 Implementation

Figure 1 displays the block diagram of the proposed model. The suggested model's operation is sequenced as follows: data collection, data processing, research into the

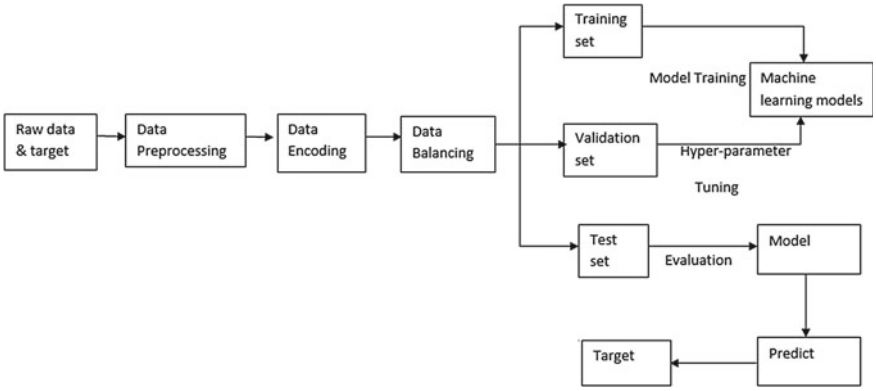

Fig. 1 Block diagram of the proposed model

appropriate model for the type of data, the model training and testing and evaluation. It is the most crucial stage in improving the accuracy of machine learning models. In supervised learning, an AI system is provided with data that has been labeled, meaning that each piece of information has been assigned to an appropriate label. Some of the most often used classification algorithms are support vector machine, naïve Bayes, logistic regression, decision trees, and KNN.

The dataset is divided into three categories as training data, validation data, and test data. To train the classifier, start with the training dataset, then fine-tune the parameters with the validation dataset, and lastly, evaluate the classifier's performance with the test dataset. The classifier has access only to the training and/or validation sets. The test dataset must not be used through classifier training. A testing set would be given mostly during the classifier's evaluation. Validation datasets are used to fine-tune the parameters of a classifier. When the categorical feature is ordinal, the categorical data encoding approach is used. The dataset is initially imbalanced, therefore, data rebalancing technique has been applied.

Figure 2 depicts the flowchart of the entire model stages followed by this technique. We have used two datasets: European and German datasets. The data is preprocessed and separated into two subsets: training and testing. The model is optimized using hyperparameter tuning, and the parameters such as accuracy, precision, recall, and F1-score are calculated using it.

4 Results and Discussions

The European credit card dataset [14] on which PCA technique has already been applied, contained 28 numerical features. The German credit card dataset [15] contained 21 features out of which 12 are categorical and 7 are numerical. The

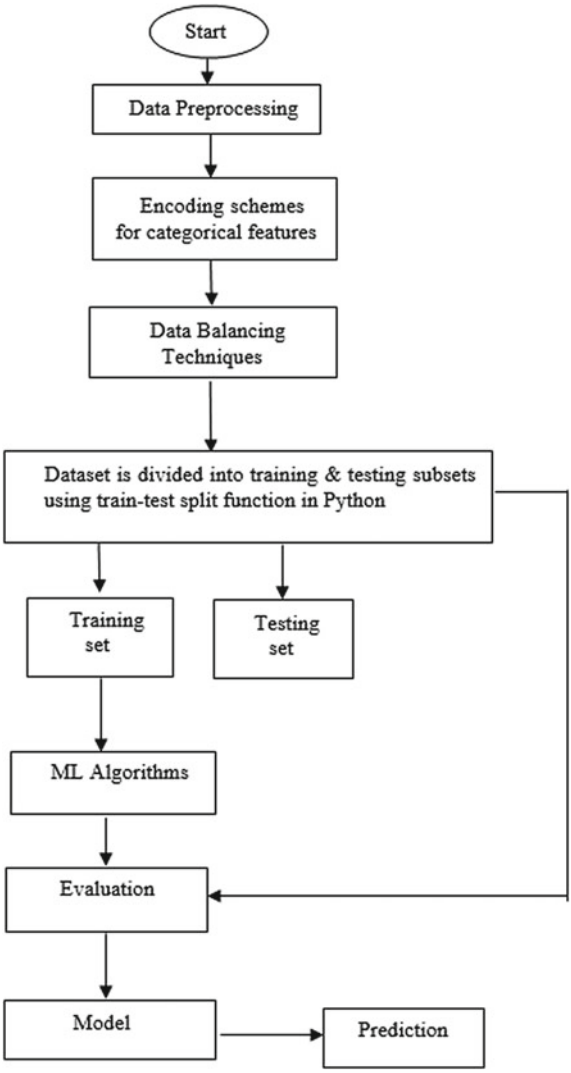

Fig. 2 Flow chart of the proposed model

proposed methodology is implemented in Python and uses machine learning classification methods. Several machine learning models such as LR, LDA, naïve Bayes, BernoulliNB, and decision tree are used to analyze it. Hyperparameter optimization is carried out using GridSearchCV (Table 3 and 4).

The results from the proposed methodology implemented using decision tree classifier are compared with that in Patil et al. [16] on German dataset which signifies the superiority of the method proposed (Table 7).

Table 1 Performance analysis without hyperparameter optimization on German dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 score
LR	71.50	69	71	0.69
LDA	69.50	66	69	0.66
Gaussian NB	71	73	72	0.73
Bernoulli NB	71.50	70	71	0.71
Decision tree	66.00	60	66	0.59

Table 2 Performance analysis without hyperparameter optimization on European dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 score
LR	99.80	100	100	1.00
LDA	98.50	97	95	0.97
Gaussian NB	91	92	93	0.92
Bernoulli NB	91.50	90	91	0.91
Decision tree	86	80	86	0.79

Table 3 Performance analysis using hyperparameter optimization on German dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 score
LR	69.50	64	69	0.64
LDA	70.7	66	69	0.66
Gaussian NB	71	73	72	0.73
Bernoulli NB	71.50	70	71	0.71
Decision tree	72.10	45	67	0.54

Table 4 Performance analysis using hyperparameter optimization on European dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 score
LR	98.5	100	100	1.00
LDA	99.6	95	97	0.95
Gaussian NB	91	93	92	0.93
Bernoulli NB	91.50	90	91	0.91
Decision tree	82.1	55	57	0.64

Table 5 Confusion matrix of decision tree classifier on German dataset

	Predicted No	Predicted Yes
Actual No	0	98
Actual Yes	0	202

Table 6 Confusion Matrix of LDA classifier on European dataset

	Predicted No	Predicted Yes
Actual No	56,852	9
Actual Yes	20	81

Table 7 Comparative analysis with the existing results

Decision tree classifier	Ref. [16]	Proposed
Accuracy	72%	72.10%

5 Conclusion

This article assesses the performance of different machine learning classification algorithms by means of a German credit card dataset to perceive whether or not an operation/transaction is fraud. The credit card dataset was imported, preprocessed, encoded, and equipped for training the model using the machine learning workflow mechanism. The models were verified using both hyperparameter optimization and non-hyper-parameter optimization methods. It was then trained, deployed, and assessed for each classification model using multiple parameters and assumptions. The decision tree classifier outperforms the LR, LDA, and naïve Bayes algorithms in terms of performance. Ensembling all utilized models utilizing voting ensemble or weighted average ensemble can help increase the model's accuracy.

References

1. Usama F, Gregory PS, Padhraic S (1996) From data mining to knowledge discovery in databases. AI Mag 17:37–54
2. Yanmin S, Wong AKC, Mohamed SK (2006) Classification of imbalanced data: a review. Int J Pattern Recognit Artif Intell 23(4):687–719
3. Şahin YG, Duman E (2011) Detecting credit card fraud by decision trees and support vector machines. In: International multiconference of engineers and computer scientists
4. Yu B, Song M, Wang L (2009) Local isolation coefficient-based outlier mining algorithm. In: International conference on information technology and computer science, vol 2. IEEE
5. Bhattacharyya S, Jha S, Tharakunnel K, Westland JC (2011) Data mining for credit card fraud: a comparative study. Decis Support Syst 50(3):602–613
6. Van Vlasselaer V, Bravo C, Caelen O, Eliassi-Rad T, Akoglu L, Snoeck M, Baesens B (2015) APATE: a novel approach for automated credit card transaction fraud detection using network-based extensions. Decis Support Syst 75:38–48
7. Yu W-F, Wang N (2009) Research on credit card fraud detection model based on distance sum. In: International joint conference on artificial intelligence. IEEE
8. Dembrani MB, Khanchandani KB, Zurani A (2017) Comparative analysis of various adaptive filter structures using Simulink. In: Proceeding of international conference on intelligent communication, control and devices. Springer, Singapore
9. Dembrani MB, Khanchandani KB, Zurani A (2017) FPGA implementation of low power & high speed accurate ECG signal detection using DABLMS algorithm. Communication and computing systems. In: Proceedings of the international conference on communication and computing systems (ICCCS 2016), Gurgaon, India
10. Guo T, Li G-Y (2007) Neural data mining for credit card fraud detection. In: International conference on machine learning and cybernetics, vol 7. IEEE
11. Panigrahi S et al (2009) Credit card fraud detection: a fusion approach using Dempster–Shafer theory and Bayesian learning. Inf Fusion 10(4):354–363
12. Srivastava A et al (2008) Credit card fraud detection using hidden Markov model. IEEE Trans Depend Secur Comput 5(1):37–48

13. Quah JTS, Sriganesh M (2008) Real-time credit card fraud detection using computational intelligence. Expert Syst Appl 35(4):1721–1732
14. Dal Pozzolo A, Caelen O, Johnson RA, Bontempi G (2015) Calibrating probability with undersampling for unbalanced classification. In: Symposium on computational intelligence and data mining (CIDM). IEEE
15. Statlog German credit card dataset, UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data. last accessed 2021/01/5
16. Patil S, Nemade V, Soni PK (2018) Predictive modelling for credit card fraud detection using data analytics. Procedia Comput Sci 132:385–395
17. Itoo F, Singh S (2021) Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. Int J Inf Technol 13(4):1503–1511
18. Thennakoon A, Bhagyani C, Premadasa S, Mihiranga S, Kuruwitaarachchi N (2019) Real-time credit card fraud detection using machine learning. In: 9th International conference on cloud computing, data science & engineering (confluence), pp 488–493
19. Cynthia PC, Thomas George S (2021) An outlier detection approach on credit card fraud detection using machine learning: a comparative analysis on supervised and unsupervised learning. In: Intelligence in big data technologies beyond the hype. Springer, Singapore, pp 125–135