Credit Card Fraud Detection

Liv Bunthorn
dept. Department of Information and Communication Engineering
Phnom Penh, Cambodia
livbunthorn@gmail.com

***Abstract*—Credit cards are the most common payment method in recent years. As technology improves, so does the number of fraud cases, ultimately requiring the development of a fraud detection algorithm to accurately detect and eliminate fraudulent activity.In this research proposes variety machine learning algorithm such as XGBoost, logistic regression, decision tree and random forest for handling imbalanced dataset.More over, in this research also study about accuracy, sensitivity, specificity, f1-score and ROC evaluation matrix to evaluation all of the algorithm to find the best algorithm. Finally, XGboost is the best algorithm compare to logistic regression, decision tree and random forest.**

***Keywords*—** Fraud detection, Credit card, Logistic regression, XGBoost, Decision tree, Random forest, accuracy, sensitivity, specificity, f1-score, ROC

I. Introduction

Credit cards are widely used [1] because of their popularity E-commerce and the development of mobile smart devices. Cardless transactions are especially popular all credit card operations are processed via internet payments such as Paypal, Bakong and Alipay. MasterCard creates web transactions simpler and more convenient. But there is an increasing trend commercial fraud causes huge monetary losses every year.Losses are expected to grow by double digits each year prices until 2020. because information about the area around the trade line and on the map is sufficient fraud is easier than ever to complete payments. tackling fraud has become a top priority for e-commerce and have a huge impact on the economy. so cheat recognition is important and necessary.

The card is only available for certain online payment methods [2]. Numbers, due dates and resumes are required and in some cases data may be lost if we are not present. We do not know how our data was stolen. Scammers use phishing techniques on the internet to get details even if we don't know our details have been compromised. To create a scam, scammer only needs some card detail for purchases, the user may not know if their balance is. Map information leaked. Card details should be kept confidential. But sometimes it is out of our control. Phishing sites can lead to information leakage, and sometimes the card itself can be lost or stolen. The best way to know if this is a business scam is to use existing data and machines to determine if a customer is making the right purchase.

The paper is organized by section II will describes of literature review about credit card fraud detection. section III will describes about proposed methodology used in the experiment. The experimentation and results will show in section IV. Finally, some conclusion is present in the last section.

II. Literature Review

The propose of this literature review is to review the previous work of researcher on detection fraud in credit card. This literature review will illustrate some recent work on the credit card fraud detection.

The term "credit card fraud" [3] describes the actual loss of a credit card or the loss of private credit card data. For detection, a variety of machine learning algorithms can be applied. Their study presents many algorithms for categorizing transactions as fraudulent or legitimate. random forest, multilayer perceptron, logistic regression and naive bayes were the algorithm employed in the experiment.

In this paper [1], train two random forests for traditional trading and abnormal trading behavior. Compare two random forests with widely different base classifiers and analyze their performance in detecting credit fraud. The information used in their experiments came from an e-commerce company in China.This author [4] mainly focus on machine learning algorithms. The algorithm program uses the Square Measure Random Forest algorithm program and the Adaboost algorithm. Result for both algorithms measure support f1-score, precision, accuracy, and recall. Plan the beast curve, backed by a confusion matrix. Comparing random forest and adaboost algorithm programs, considering the algorithm with the best precision, accuracy, F1-score, and recall as the best algorithm program for detecting scams. Tend [5] to take the utilization of prognostic Associate in Nursing Altaic's done by the enforced machine learning models and an API module to come to a decision if a selected dealings is real or fallacious. we tend to additionally assess a unique strategy that effectively addresses the skew distribution of information. the info utilized in our experiments come back from a financial organization in line with a confidential revelation agreement.

Scammers [6] identify opportunities to steal users' Mastercard data through deceptive text messages and phone calls, as well as camouflage attacks, phishing attacks, and more. When manipulating various machine learning algorithms such as support vector machines (SVM), k-nearest neighbors (Knn), and artificial neural networks (ANN) to predict the occurrence of fraud. Furthermore, we tend to distinguish between supervised machine learning and deep learning performed.An introduction [7] to the fraud structures associated with credit cards and their different types. Explains various techniques that can be used in system detect fraud such as Artificial Neural Networks (ANN),Support Vector Machines (SVM), Bayesian Networks, K-Nearest Neighbors (KNN), Hidden Markov Models, systems based mainly on and Symbolic Logic for Call Trees. Completed a comprehensive review of current and planned Mastercard fraud detection models and conducted a comparative study of these techniques related to the idea of quantitative measures such as accuracy, detection rate, and warning rate. Our findings explain the shortcomings of existing models and provide an improved answer to beat them.

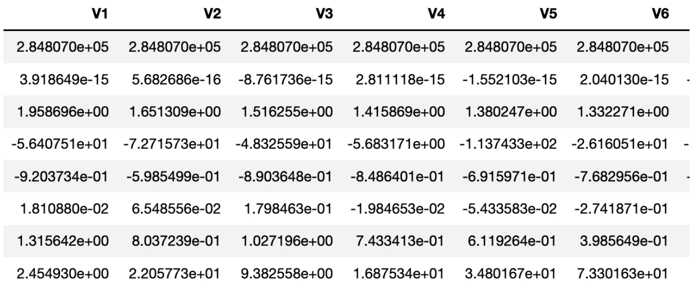
The combination of data mining techniques and machine learning [8] were ready to establish the transactions by learning the patterns of information. once preprocessing with the dataset normalization and victimization principal element analysis all the classification achieve more than 95 percentages accuracy compare to result earned before preprocessing on the dataset.Various approaches [9] are found by several researchers until date to find these frauds and to scale back them.The comparison of native isolation issue and outlier issue algorithms mistreatment and their detailed experiment result area. once the analysis of the dataset their have a tendency to got 97% on native outlier issue and 76% on isolation forest.Fraud protection [10] system, currently become necessary to remove the losses of banks and money establishments.Throughout this analysis article,have a tendency to introduce a good credit card fraud detection system together with a feedback mechanism.German credit card dataset [11] is employed to judge these machine learning algorithm potency supported filter and wrapper options choice technique. The experimentation outcome show about the accuracy of J48 and half has been hyperbolic when apply filter and wrapper strategies. Finally, preciseness and sensitivity of J48, AdaBoost, and therefore the random forest are increased.

III. Proposed Methodology

The dataset get for European cardholders in September 2013. Dataset represents transaction that occur two days, and out of 284,807 transactions, we have 492 cases of fraud. The dataset is very imbalanced, with the positive class (fraud) accounting for 0.172 percentages of all transactions.

This is all the dataset when we do a virtaulization.

Since all the columns of this dataset already PCA transformed,So we do not need to do data preprocessing and outlier detection.



V1	V2	V3	V4	V5	V6
2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05
3.918649e-15	5.682686e-16	-8.761736e-15	2.811118e-15	-1.552103e-15	2.040130e-15
1.958696e+00	1.651309e+00	1.516255e+00	1.415869e+00	1.380247e+00	1.332271e+00
-5.640751e+01	-7.271573e+01	-4.832559e+01	-5.683171e+00	-1.137433e+02	-2.616051e+01
-9.203734e-01	-5.985499e-01	-8.903648e-01	-8.486401e-01	-6.915971e-01	-7.682956e-01
1.810880e-02	6.548556e-02	1.798463e-01	-1.984653e-02	-5.433583e-02	-2.741871e-01
1.315642e+00	8.037239e-01	1.027196e+00	7.433413e-01	6.119264e-01	3.985649e-01
2.454930e+00	2.205773e+01	9.382558e+00	1.687534e+01	3.480167e+01	7.330163e+01

Fig. 1: dataset of v1-v6

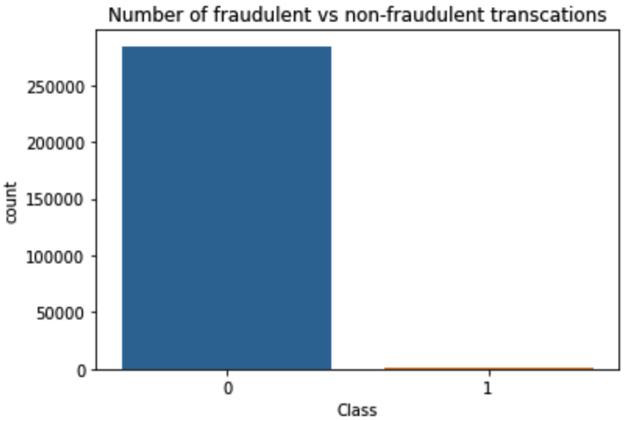


Fig. 2: virtualization of all dataset

All of the dataset is divide to 20% for testing and 80% for training set. Now we are starting to implement our algorithms.

Logistic Regression : is a estimates the probability of an event occurring, for example, to choose or not to choose based on a given set of independent variables. Since the outcomes are probabilities, the dependent variable ranges between 0 and 1.Logistic regression applies a logit transformation to odds, which is the probability of success divided by the probability of failure. This is also commonly referred to as log odds or the natural logarithm of odds, and this logistic function is represented by the following formula:

$$logit(P) = \ln \frac{(P)}{(1-P)}$$

XGBoost: is a classifier algorithm of Machine learning that's applied for structured and tabular information. XGBoost is associate implementation of gradient boosted call trees designed for speed and performance. XGBoost is associate extreme gradient boost rule. which means that it's an enormous Machine learning rule with countless components. XGBoost works with massive, sophisticated datasets. XGBoost is associate ensemble modelling technique.

Decision trees: is a supervised learning algorithm, this algorithm can apply on both regression and classification. Decision tree performance better with classification problem. Decision algorithm is a tree structure classifier, use internal node for the feature of the dataset. During implement this algorithm the major problem is about choose a good attributes for the and child nodes. To address this problem, we can use sttribute selection measure technique.Two popular technique for attribute selection measure are:

- Information Gain(IG):

$$IG = Entropy(s) - [(WeightedAvg) \times Entropy(eachfeature)]$$

where,

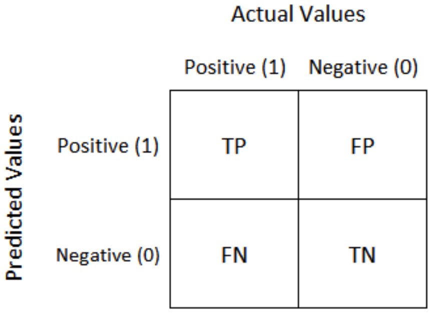
$$Entropy(s) = -P(yes) \times \log 2 \times P(yes) - P(no) \times \log 2 \times P(no)$$

- Gini Index:

$$GiniIndex = 1 - \Sigma_j P_j^2$$

Random forest: is a algorithm in machine learning this technique from supervised learning.This algorithm can apply on both regression and classification.The algorithm of random forest is build decision trees on variety samples and base on the major vote for classification and use average for regression. This algorithm also a good algorithm to handle with imbalance dataset.

Confusion Matrix: is table that use to show the performance of classification model. This table is very importance to measure accuracy, sensitivity, specificity, f1-score and ROC.. Fig.3 shows a confusion matrix table.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Fig. 3: Confusion Matrix table

Accuracy: is the correction of predicting output in percentage.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)}$$

Sensitivity :is refers to the probability of a positive test and condition on actually is positive.

$$Sensitivity = \frac{TP}{(TP + FN)}$$

Specificity :refers to the probability of a negative test and the conditioned on actually is negative.

$$Specificity = \frac{TN}{(TN + FP)}$$

F1-score : is the average of precision and recall.

$$F1 - score = \frac{2 \times (Recall \times Precision)}{(Recall + Precision)}$$

$$Recall = \frac{TP}{(TP + FN)}$$

$$Precision = \frac{TP}{(TP + FP)}$$

ROC curve or Receiver Operating Characteristic Curve: is a graph show about the performance of classifier model.The plot of ROC curve have two parameters:

- True Positive Rate (Y-axis)

$$TruePositiveRate(TPR) = \frac{TP}{(TP + FN)}$$

- False Positive Rate (X-axis)

$$FalsePositiveRate(FPR) = \frac{FP}{(FP + TN)}$$

IV. Experimentation and Results

In this section I will compare all the results of the experiment and find the best algorithm in this project.

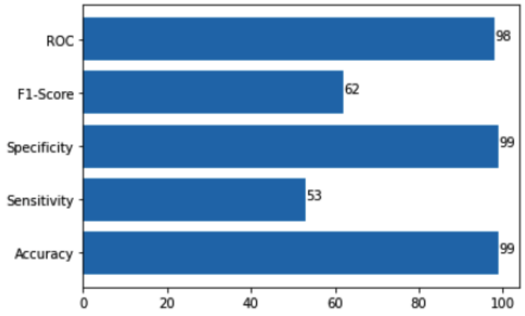


Fig. 4: Result of Logistic Regression

As we can see in the Fig.4 Logistic Algorithm is good results with ROC, Specificity and Accuracy evaluation matrix, but this algorithm not suitable with F1-Score and Sensitivity evaluation matrix.

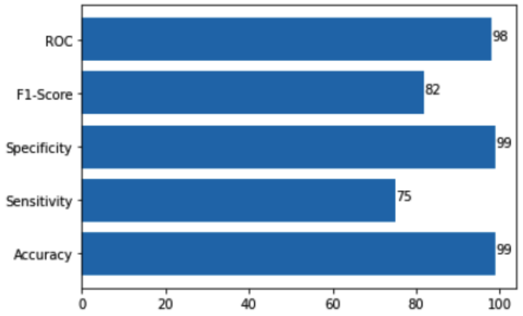


Fig. 5: Result of XGBoost

Fig.5 is the result of XGBoost, we can see XGBoost is good performance for all evaluation matrix. This algorithm got 98% for ROC, 82% for f1-Score, 99% for specificity, 75% for sensitivity and 99% for accuracy matrix.

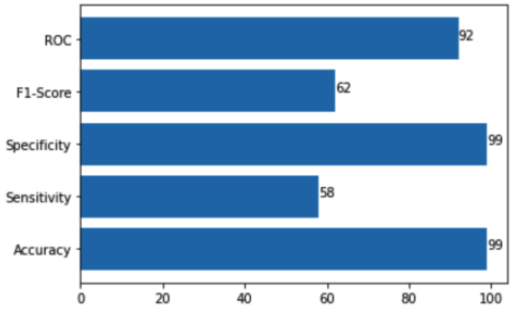


Fig. 6: Result of Decision Tree

Decision Tree algorithm got only 62% with f1-score evaluation matrix and 58% with sensitivity evaluation matrix, In contract Decision Tree algorithm good performance with ROC, specificity and accuracy.

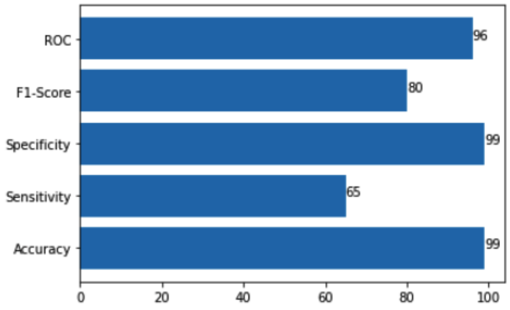


Fig. 7: Result of Random Forest

Random Forest in the last algorithm in this experiment,as we can see the result in Fig. 7 Random Forest got more than 95% with ROC, spesificity and accuracy evaluation matrix, while this algorithm got 80% with f1-score and only 65% with sensitivity evaluation matrix.

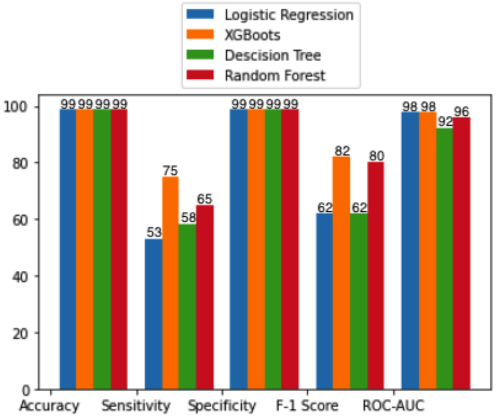


Fig. 8: Result of all algorithms

According the Fig. 8 among all the algorithm we can see XGBoost is the best algorithm compare to logistic regression, decision tree and random forest.

V. Conclusion

In this paper we studied about algorithm logistic regression, XGBoost, decision tree and random forest.after experimented on all the algorithm, we can see logistic regression, decision tree and random forest got over 95% on accuracy, specificity and ROC evaluation matrix, but got less than 65% on sensitivity evaluation matrix. In f1-score evaluation matrix logistic regression algorithm got 62%, decision tree algorithm got 62% random forest got 80%.In experiment of XGBoost algorithm got 98% ROC, 82% f1-score, 99% specificity, 75% sensitivity and 99% accuracy. In comparison, XGBoost is the best algorithm compare to logistic regression, decision tree and random forest.

References

[1] Shiyang Xuan, Guanjun Liu, Zhenchuan Li, Lutao Zheng, Shuo Wang, and Changjun Jiang. Random forest for credit card fraud detection. In *2018 IEEE 15th international conference on networking, sensing and control (ICNSC)*, pages 1–6. IEEE, 2018.

[2] D Tanouz, R Raja Subramanian, D Eswar, GV Parameswara Reddy, A Ranjith Kumar, and CH VNM Praneeth. Credit card fraud detection using machine learning. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 967–972. IEEE, 2021.

[3] Dejan Varmedja, Mirjana Karanovic, Srdjan Sladojevic, Marko Arsenovic, and Andras Anderla. Credit card fraud detection-machine learning methods. In *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)*, pages 1–5. IEEE, 2019.

[4] Ruttala Sailusha, V Gnaneswar, R Ramesh, and G Ramakoteswara Rao. Credit card fraud detection using machine learning. In *2020 4th international conference on intelligent computing and control systems (ICICCS)*, pages 1264–1270. IEEE, 2020.

[5] Anuruddha Thennakoon, Chee Bhagyani, Sasitha Premadasa, Shalitha Mihiranga, and Nuwan Kuruwitaarachchi. Real-time credit card fraud detection using machine learning. In *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 488–493. IEEE, 2019.

[6] RB Asha and Suresh Kumar KR. Credit card fraud detection using artificial neural network. *Global Transitions Proceedings*, 2(1):35–41, 2021.

[7] Yashvi Jain, Namrata Tiwari, Shripriya Dubey, and Sarika Jain. A comparative analysis of various credit card fraud detection techniques. *Int J Recent Technol Eng*, 7(5S2):402–407, 2019.

[8] Ong Shu Yee, Saravanan Sagadevan, and NHAH Malim. Credit card fraud detection using machine learning as data mining technique. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, 10(1-4):23–27, 2018.

[9] Hyder John and Sameena Naaz. Credit card fraud detection using local outlier factor and isolation forest. *Int. J. Comput. Sci. Eng*, 7(4):1060–1064, 2019.

[10] Naresh Kumar Trivedi, Sarita Simaiya, Umesh Kumar Lilhore, and Sanjeev Kumar Sharma. An efficient credit card fraud detection model based on machine learning methods. *International Journal of Advanced Science and Technology*, 29(5):3414–3424, 2020.

[11] Ajeet Singh and Anurag Jain. Adaptive credit card fraud detection techniques based on feature selection method. In *Advances in computer communication and computational sciences*, pages 167–178. Springer, 2019.