JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

MACHINE-LEARNING ENHANCED FRAUD PREVENTION

Priya Pai
E&CE, PESITM, Shivamogga

Preethi G Y
E&CE, PESITM, Shivamogga

Dr. Vishnu V M
Associate Professor
E&CE, PESITM, Shivamogga.

Ranjitha G
E&CE, PESITM, Shivamogga

Mehnaz Fathima
E&CE, PESITM, Shivamogga

Abstract:

Fraud involving credit cards is becoming a major worry for both people and financial organizations as the world of online financial transactions continues to grow at an incredible rate. Our research presents a technique which helps to determine the fraud caused in the use of credit cards. By utilizing combination of both under supervision and algorithms, our suggested system can quickly and accurately identify any fraudulent activity that occurs in the present. We assess the system's performance using important measures like accuracy, and an F1 score on a dataset containing both authentic and illicit transactions to insure that gauge its efficacy. The algorithms used for identifying credit card that are fraudulent are Logistic Regression, Random Forest, k-nearest neighbor (kNN), Support Vector Machine(SVM). Remarkably, about 70% of the people acknowledged having at least one credit card. Regrettably, utilising credit cards has grown in tandem with credit cards' growing usage. There are two sorts of these fraudulent acts. In the first, An ATM card account is opened in the appellation of the victim by an identity thief; statistics indicate that this type of fraud increased by a staggering 48% between 2019 and 2020. The second sort of deception, which accounted for 9% more reports in 2019 and 2020, occurs when an identity thief gains having entry to a person active credit card accounts. The startling growth in credit card theft underscores the requirement for additional security measures to safeguard consumers' financial security. Examining the exploratory work, the findings indicate that the ROC curves with optimized values, perfection, F1-score, and delicacy have all significantly improved. This inspired to research various machine learning methods and approaches in order to address the issue of identifying credit card purchases that are fraudulent.

Keywords: Using Artificial Intelligence to Increase Security. Logistic Regression, Random Forest, k-nearest neighbor (kNN), Support Vector Machine (SVM). For the best financial security, measure success using ROC curve, F1 score, recall, and precision.

Introduction:

The incidence of fiscal the amount of fraud has gone up recently, primarily as a result of technological developments and the growth of sectors like ecommerce and fiscal technology (also known as Fin Tech). Because of these developments in technology, there acted as a rise in credit card transactions, which has given rise to fraud. More particular, there acted as a notable increase in the quantity of incidents involving credit card theft. When someone improperly and without authorization uses another person's credit card, it is referred to as credit card theft. This may occur if the offender uses one of several fraudulent methods in order to obtain the credit card information, including intercepting a valid transaction or cloning the actual card. Moreover, a wide range of organizations, including card issuers, retailers, and small companies, are significantly impacted by credit card theft. As per Khin Wee Lai, the associate editor of this newspaper, credit card theft cost the world an astounding \$21.84 billion for the year 2015. This figure kept going up, with losses in 2019 amounting to \$28.65 billion, a four-fold increase of \$6.81 billion. In 2020, out of almost 1.4 million occurrences of identity theft, 393,207 instances is referred to as credit card theft were reported. This trend persisted. Right present, Fraud involving credit cards is the second most common type of identity theft, only surpassed by benefits and government document fraud. The previous year, theft of credit cards cost the world economy a staggering $24.26 billion. Based on recent data, the US is the most vulnerable nation to this kind of crime, accounting for 38.6% of all reported losses resulting from credit card theft in 2018. Implementing techniques to identify credit card theft that can successfully protect the dependability and security of all elements involved using credit card dealings is therefore imperative. The dataset presents a challenge to conventional approaches due of its extreme imbalance. We address this by using novel method that combines highly optimized techniques like logistic regression, Random Forest, k-nearest neighbour (kNN), and the SVM (support vector machine) with readily available datasets. A credit card theft identifying system's ultimate purpose is to accurately identify a greater number of fraudulent cases to be able to foster client trust in the financial institution and avoid financial losses brought on by false positives. This study presents a scalable methodology in order to recognise a credit card fraud and try to further this objective. To verify the efficacy of the suggested framework, it is evaluated on a synthetic dataset that is noticeably unbalanced.

The remainder of this essay is set up as follows: We go into earlier research on machine learning's application to identify the credit card fraud in Section II. Our thorough a plan in place to recognise a credit card fraud is presented in Section III, together with specifics on the dataset, pre-processing strategies, feature extraction approaches, chosen algorithms, our framework, and the assessment criteria that were employed. In Section IV, we proceed to analyse the outcomes of our investigations. In Section V, we conclude our paper and offer some recommendations for further investigation.

II. Literature Review:

Scholars have put out a few of strategies to stop fraudulent transactions and spot credit card fraud. An overview of recently discovered cutting-edge approaches is provided in this article. The AIS- rested fraud discovery model (AFDM), a unique technique, was introduced by Halvaiee and Akbari. Through the execution of the Immune System Inspired Algorithm (AIRS), they have effectively improved the fraud detection accuracy. Their research shows that, in comparison to conventional algorithms, AFDM enhances detection rates by more than 25%, lowers expenses by more than 85%, and shortens system reaction times by more than 40%[1]. Khatri et al. [2] tested many machine learning algorithms in their study in order to identify fraudulent credit card transactions. They employed well known methods like Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), and k-Nearest Neighbour (kNN), as well as Naive Bayes (NB). After collecting data from European cardholders in 2013[5], the authors evaluated the measures for performance for precision and tact to evaluate the efficacy of this machine learning (ML)-based models detected fraud.

The outcomes demonstrated that the kNN algorithm performed superior to any other, obtaining a remarkably high sensitivity of 81.19% and accuracy of 91.11%. Expanding upon this study, Rajora et al. [4] investigated the utilisation of numerous methods of credit card fraud detection using machine learning, with a focus in particular on the European market. The authors determined that the two most crucial actions for assessing performance were the delicacy and area under the wind (AUC). According to their investigation, the kNN technique produced an AUC of 0.94 and an exactness of 93.2%, little less substantial than the RF algorithm's remarkable 94.9% accuracy. Although the outcomes are promising, the disparity in class in the dataset that the researchers employed was not addressed by them. Using the European cardholder dataset, the researchers investigated many techniques for machine learning under supervision which includes Random Forest (RF) and Gradient Boosting (GB).

Sensitivity and accuracy measures were employed to assess the efficacy of various methods. As stated by the experiment results, GB obtained a precision and accuracy of 93.99% and 94.01%, respectively, while RF obtained a precision and accuracy of 95.98% and 94.00%, respectively [11]. Tanouz et al. [12] presented a technique in order to identify credit card fraud based on machine learning this research. Within this research, the researchers utilized the European card holders' dataset to assess the efficacy of their proposed methods. Additionally, they implemented an under sampling technique to deal with the dataset's imbalanced class problem. The recognition of fraudulent instances was the primary measure of performance that was employed, or sensitivity. The findings showed that the Random Forest (RF) approach achieved a sensitivity of 91.24, while the Logistic Regression (LR) model achieved a sensitivity of 95.16.

This study looks at how different styles—Random Forest (RF), k-nearest neighbor (kNN), Support vector machine(SVM) and logistic retrogression(LR)—are utilized the given job. This study's main goal is to use these techniques to stop credit card theft. The Random Forest model produced an excellent accuracy of 94.9%, while the kNN model acquired a sensitivity of 93.2%, according to the findings. Specially, these findings suggest that there is room for maximum accuracy in this field.

III. Proposed approach to detecting credit card fraud:

The proposed frame for fraud discovery is presented in Fig 1. As this image illustrates, we begin by applying the requested pre-processing to the data and then divide it into training and testing portions. Next, we use Synthetic Minority Oversampling Technique (SMOTE) on the training data to identify the fashionable hyperparameters that improve performance. In addition to examining the algorithms using an assortment of assessment criteria, including as delicacy, recall, the Matthews correlation measure (MCC), the F1-score, and ROC Curve, we apply the cross-validation technique to provide performance comparison in an unstable set. The following is a detailed explanation of these ways:

A. DATASET:

This task shows how our suggested algorithm could be involved with practice using an actual dataset. The 'creditcard' dataset includes 492 examples of dishonest business dealings mixed in with the bulk of legitimate transactions. The prevalence rate of 0.172, indicates the proportion of transactions that are fraudulent in the dataset, makes the extreme imbalance obvious in the dataset. You may get this dataset by going to https://www.kaggle.com/mlgub/creditcardfraud.

B. DATA PRE-PROCESSING:

The large disparity between the quantity of phony and real transactions points to an unsteady data distribution. In real-world fraudulent use of credit cards datasets, this frequent occurrence could have a detrimental effect on the performance of machine learning algorithms. Actually, adding a class with little representation can seriously distort the result of the evaluation. Numerous studies have attempted to balance the data by using undersampling or oversampling approaches in an effort to address this problem. Conversely, however, oversampling might provide duplicate information that muddies the data, while undersampling might result in the demise of important data. In this experiment have used the synthetic minority oversampling method (SMOTE) as a potential remedy.

C. FEATURE EXTRACTION:

The time (in seconds) that elapsed between the initial sale and each subsequent sale is included into the "time" point. In order to maximise the point, we extend it to include the sale hour point, which provides us with more details than the time point alone.

D. FEATURE SELECTION:

In the provided data, "time" and "quantum" certain components that have been certain components that have been recognized; no new information has been added. Point selection is used to find a subset of characteristics in order to be capable to enhance the classifier's performance in correctly identifying credit card fraud. The training data's dimensions are decreased and the salient characteristics are determined using the information gain (IG) system. The way this approach operates is by highlighting the commonalities in credit card transactions and giving greater weight to the characteristics that most effectively differentiate between authentic and fraudulent transactions. The efficiency of the IG system has proven, in addition to its exceptional accuracy performance. Because of this, our study also takes the IG system into account.

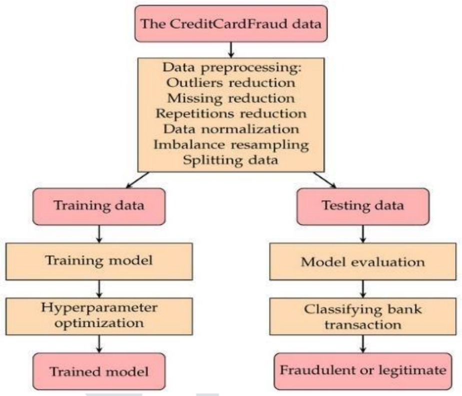

Figure 1.Proposed methodology for identifying credit card theft.

E. ALGORITHMS:

It is impossible to overestimate the influence of hyperparameters on the carrying out of machine learning models. Optimization is the method by which modifying the optimal hyperparameter values during algorithm for machine learning training. When compared to conventional evolutionary strategies, recent research has demonstrated that Bayesian optimization techniques are remarkably effective at determining the best values with fewer training iterations. In this work, we use SMOTE method to tune our hyperparameters, which results in improved performance and shorter computation times.

1. LOGISTIC REGRESSION

Binary outcome prediction is a secured application of numerical methods like logistic regression. This technique is specifically utilized in the domain detecting credit card theft in order to categorise transactions as either fraudulent (class 1) or non-fraudulent (class 0) depending on a few of input variables. The method determines which class an input is most likely to belong to by using a logistic function. It then allocates the input to the class having the greatest probability.

2. SUPPORT VECTOR MACHINES (SVM)

SVM is a vital machine learning method that's widely applied to tasks including regression and classification. This is accomplished by determining the best separation hyperplane between data points from various classifications. SVM can efficiently generate a decision boundary that optimizes the distinction between ATM card theft detection: transactions that are fraudulent and those that are not. Making advantage of different kernel functions, it can handle information that is both linearly and non-linearly separable, demonstrating its versatility.

3. K-NEAREST NEIGHBORS (KNN)

The k-nearest neighbors algorithm, or KNN, is a useful tool for regression and bracketing applications. Essentially, it groups fresh examples in a space based on their proximity to the k closest training cases. This method, which entails determining how far away a sale is from its k nearest neighbors, can be quite helpful in spotting credit card fraud. A voting procedure as per the agreement about the sales nearby cases then determines the sale's final classification. KNN is a dependable and effective algorithm in a range of circumstances, such as identification of ATM card frauds, thanks to its uncomplicated methodology and useful applications.

4. RANDOM FOREST

Several decision trees are worked by the potent tool Random Forest to produce precise forecasts. A Random Forest model is evolved in the area predicting credit card theft by building several decision trees on arbitrary selections of the data. Without any guidance, a new collection of features is selected at every split. The outcomes of each individual decision tree are combined to get the final prediction, usually via voting or average. By using this method, Random Forest is capable of handling complicated feature interactions and avoid overfitting.

F. EVALUATION METRICS

Specifically, we employ a cross-validation test to determine how effectively our proposed model detects credit card fraud. By using a stratified 5-fold confirmation test, this technique guarantees accurate performance comparisons even in unstable sets. Five equal-sized from our dataset, subsets are chosen at random, and the quantity of samples in each class is spread equally within each subset. One subset (20% of the dataset) is reserved for verifying the effectiveness of our suggested method for each confirmation cycle, with the remaining four subsets (80% of the dataset) being utilized for training. This procedure is carried out five times, using all subsets until they have all undergone extensive testing. The final result comes from a 5-fold cross-validation test that assesses the suggested technique and is based upon the average performance of the five test subgroups. We follow the standard criteria for our evaluations, which include sensitivity, accuracy, recall, the Matthews correlation coefficient (MCC), the F1-score, and ROC Curve, to guarantee objectivity in our comparisons. Positive values in our trials represent fraudulent transactions, and negative values represent real ones. False positives (FP) indicate the quantity of normal transactions that are incorrectly identified as fraudulent, whereas true positives (TP) indicate the number of false transactions that occur correctly identified. However, dishonest dealings that are inadvertently mislabelled as real are referred to as false negatives (FN), legitimate transactions are classified as true negatives (TN). These criteria are accurately described as Eq. (1) through Eq. (5).

Accuracy = TP+TN TP+TN+FP+FN (1)

Recall = TP TP+TN (2)

Precision=TP TP+TN+FP+FN (3)

F1-Score=2×Precision×RecallPrecisionRecall (4)

MCC=TP×FP-FP×FN√(TP+FP)(TP+FN)(TN+FP)(TN+FN) (5)

The percentage of cases accurately classified out of all cases is called accuracy. Recall measures the percentage of real positives relative to all positive cases. Precision computes the percentage of real positives relative to all cases with positive predictions. F1-Score, which provides a fair assessment of both recall and precision. The Matthews Correlation Coefficient, or MCC, evaluates overall performance by accounting for both true and erroneous positives and negatives. The accuracy metric, which counts the count of correct predictions made, represents the classifier's overall proficiency. But this standard might also lead to inadequate outcomes when addressing unbalanced data because it emphasizes a single fraudulent incident. Conversely, recall demonstrates how well the classifier identified actual fraudulent transactions. The F1-Score, however, in contrast, balances recall and accuracy by keeping in mind both incorrect negatives and positives. Accuracy, instead, reveals the classifier's dependability. Furthermore, the ROC-AUC metric assesses the model's separability by evaluating its capacity to distinguish between classes. The rates of true positives (TPR) and false positive rate (FPR) have the following relationship represented visually by the ROC-AUC, which serves as a visual depiction of the accuracy of methods for detecting fraud. But because it only takes into account positives, it is unable to appropriately assess a classifier's effectiveness. To ensure that facilitate classifier comparisons, measures of recall and accuracy are frequently employed. In this two dimensional graph, the recall is usually displayed on the x-axis and the perfection rate on the y-axis.

	Model	Accuracy	ROC Score	F1 Score	Precision Score	Recall Score
1	Logistic Regression	0.943	0.8360	0.090	0.048	0.727
2	SVM	0.871	0.8561	0.165	0.093	0.740
3	KNN	0.972	0.8171	0.043	0.022	0.761
4	Random Forest	0.984	0.9134	0.289	0.174	0.842

Regretfully, no one index can show the real and false positives and negatives sufficiently. The MCC is a helpful for evaluating a two-class problem since it considers both true and false positives and negatives. In fact, given the different sizes of the courses, it's a well-balanced measure.

IV. Outcome and Discussion

Boosting algorithms with SMOTE method and the stratified 5-fold Techniques for cross-validation are utilised to evaluate the efficiency of our suggested framework. The algorithms are assessed separately and their hyperparameters are taken out prior to using the majority voting technique. We perform tests in both triple and double precision; Table 1 shows the comparison outcomes.

Table 1. Comparison results of models

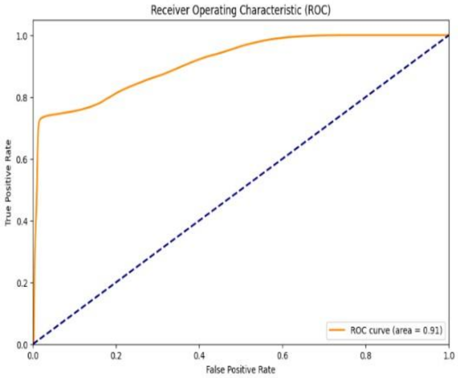

Figure 2.ROC curve of logistic regression

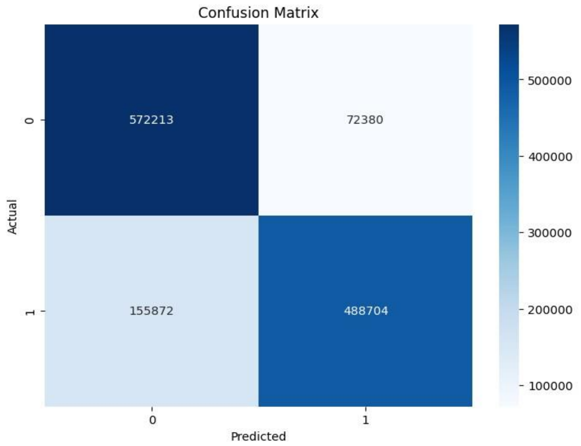

Figure 3.Confusion matrix of logistic regression

Figures depicting the assessment of each individual's performance models reveal the ROC curve and confusion matrix, both of which offer valuable insights. The ROC curve showcases the true positive and false positive rates, while the confusion matrix displays the actual and predicted values. In particular, the ROC curve of figure 2 exhibits an impressive area under the curve of 0.83 for logistic regression, while figure 3 presents a corresponding confusion matrix with an accuracy of 81.79%, the ROC curve of figure 4 exhibits an impressive area under the curve of 0.85 for SVM, while figure 5 presents a corresponding confusion matrix with an accuracy of 86.08%.

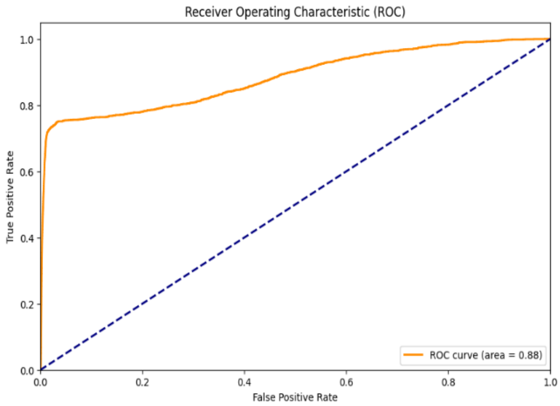

Figure 4.ROC curve of SVM

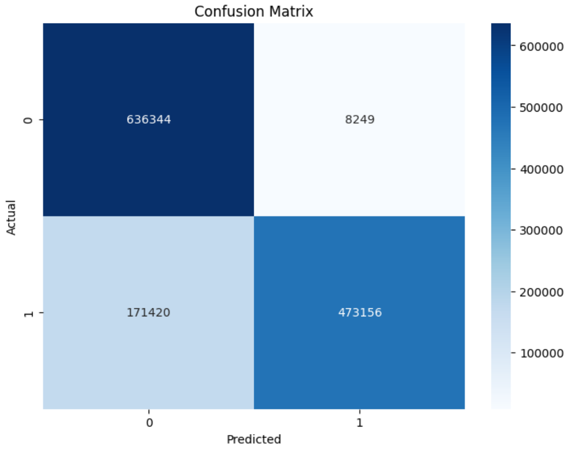

Figure 5.Confusion matrix of SVM

For an illustration of fraudulent or legitimate transactions, various features such as transaction time, gender, location (latitude and longitude), credit card number, merchant, category, transaction amount, first and last names, etc. are considered. Additionally, figure 6 displays the distribution of transaction amounts, while the distribution of gender is as picturised in figure 7. Figure 8 displays the transaction time and fraud correlation, while the distribution of fraudulent transactions is as picturised in figure 9.

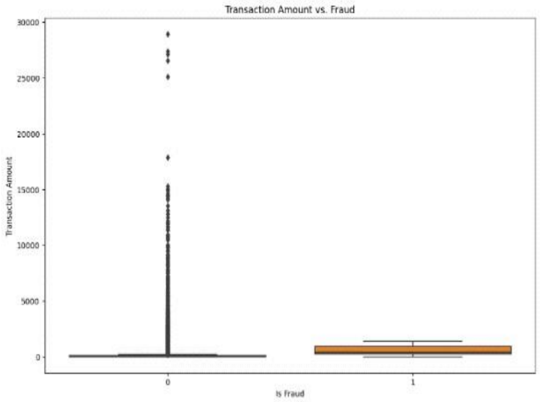

Figure 6.Distribution of transaction amounts

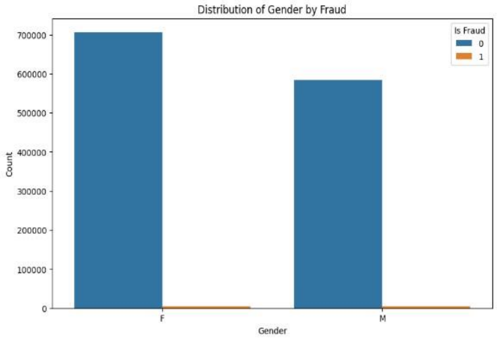

Figure 7.Distribution of gender

V. Conclusion and Future Work:

The challenge of identifying real-world credit card fraud datasets with unbalanced data was the main emphasis of this study. We used to solve this issue, a machine learning method to raise the precision of fraud detection. The examination of publicly accessible dataset pertaining solely to "credit cards" was the foundation of our research. In our investigation, we considered the following four methods: SVM, KNN, Random Forest, and Logistic Regression. The techniques were assessed using commonly used measures like as accuracy, precision, recall method, F1-score, and ROC-AUC.

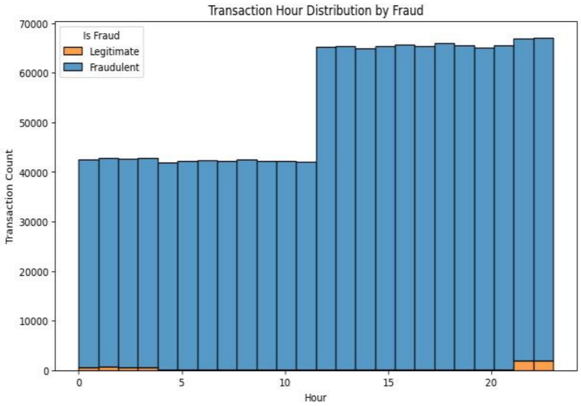

Figure 8.Transaction time and fraud correlation

Our findings demonstrated that the Random Forest approach outperformed the recently disclosed methodology, raising the F1-score by 4.9% and the exactness of false transaction identification by 94.16%. Furthermore, we implemented the majority voting algorithm and can successfully increase the algorithm's performance. MCC's performance on unbalanced data has proven to be superior to other evaluation criteria. This study accomplishes two goals: first, it conducts exploratory data analysis to find useful insights and develop a better comprehension of the dataset; second, it searches for potential connections between different variables. With regard to further study, we can build on these goals through the usage of algorithms in machine learning methods to produce predictions models and assessing and contrasting their performance to select the finest one. Additionally, using more sophisticated technology in subsequent research can provide even greater advancements, which might then be analysed with the results of this study.

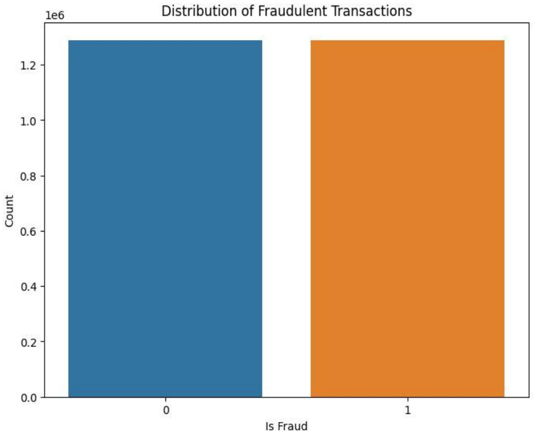

Figure 9. Distribution of fraudulent transactions

VI. References

[1] N. S. Halvaiee and M. K. Akbari, "A novel model for credit card fraud detection using artificial immune systems," Appl. Soft Comput., vol. 24, pp. 40–49, Nov. 2014.

[2] S. Khatri, A. Arora, and A. P. Agrawal, "Supervised machine learning algorithms for credit card fraud detection: A comparison," in Proc. 10th Int. Conf. Cloud Comput., Data Sci. Eng. (Confluence), Jan. 2020, pp. 680–683.

[3] A. Rb and S. K. Kr, "Artificial neural network detection of credit card fraud,", vol.2, Jun. 2021.

[4] S. Rajora, D. L. Li, C. Jha, N. Bharill, O. P. Patel, S. Joshi, D. Puthal, and M. Prasad, "A comparative study of machine learning techniques for credit card [10] N. K. Trivedi, S. Simaiya, U. K. Lilhore, and S. K. Sharma, "An efficient credit card fraud detection model based on machine learning methods," Int. J. Adv. Sci. Technol., vol. 29, no. 5, pp. 3414–3424, 2020.

[11] A. Tharwat, "Parameter investigation of support vector machine classifier with kernel functions," Knowl. Inf. Syst., vol. 61, no. 3, pp. 1269–1302, Dec. 2019.

[12] R. Sailusha, V. Gnaneswar, R. Ramesh, and G. R. Rao, "Credit card fraud detection using machine learning," in Proc. 4th Int. Conf. Intell. Comput. Control Syst. (ICICCS), May 2020, pp. 967–972.

[13] K. Kirasich, T. Smith, and B. Sadler, "Random forest vs logistic regression: Binary classification for heterogeneous datasets," SMU Data Sci. Rev., vol. 1, no. 3, p. 9, 2018.

[14] T. Hengl, M. Nussbaum, M. N. Wright, G. B. M. Heuvelink, and B. Gräler, "Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables," PeerJ, vol. 6, p. e5518, Aug. 2018.

[15] J. Feng, H. Xu, S. Mannor, and S. Yan, "Robust logistic regression and classification," in Proc. Adv. Neural Inf. Process. Syst., vol. 27, 2014, pp. 253–261.

fraud detection based on time variance," in Proc. IEEE Symp. Comput. Intell. (SSCI), Nov. 2018, pp. 1958–1963.

[5] Credit Card Fraud Detection. Accessed: Sep. 27, 2021.[Online].Available:https://www.kaggle.com/mlg-ulb/creditcardfraud

[6] Google Colab. Accessed: Sep. 27, 2021. [Online]. Available: https://colab.research.google.com/

[7]Scikit-learn: Machine Learning in Python. Accessed:Sep.27,2021.[Online].Available:https://scikit-learn.org/stable/

[8] Pandas. Accessed: Sep. 27, 2021. [Online]. Available: https://pandas. pydata.org/

[9] S. Khatri, A. Arora, and A. P. Agrawal, "Supervised machine learning algorithms for credit card fraud detection: A comparison," in Proc. 10th Int. Conf. Cloud Comput., Data Sci. Eng. (Confluence), Jan. 2020, pp. 680–683