Utilization of Machine Learning Strategies in the Investigation of Suspected Credit Card Fraud

Rifat Al Mamun Rudro
Department of Computer Science and Engineering, American International University, Bangladesh
Email: rifat.rudro138964@gmail.com

Md. Faruk Abdullah Al Sohan
Department of Computer Science, American International University, Bangladesh
Email : faruk.sohan@aiub.edu

ABSTRACT

Credit card fraud transactions have been one of the most difficult issues for banks and other financial institutions in recent years. In such events, billions of dollars are lost by financial institutions and the banking system. Concurrently, user information is not safe for that purpose. To address these issues, this paper proposes an efficient solution to automate the task using machine learning techniques such as SMOTE and ADASYN. This paper also intends to run machine learning supervised models. We discovered class imbalancing issues after examining the experiment outcomes on European cardholder datasets. Oversampling and under sampling strategies are utilized to solve fraud situations to avoid them. Predictive models such as the LR, K-nearest neighbors, decision tree, random forest XGBoost, and support vector machines are utilized to achieve the model accuracy required to find the most fit-able models for credit card fraud. The performance of SMOTE machine learning approaches increased with a 0.96 model accuracy in random forest and XGBoost.

Keywords – **SMOTE, ADASYN, XGBoost, Machine Learning, SVM, Precision, Recall**

Date of Submission: June 12, 2023 Date of Acceptance: August 06, 2023

I. INTRODUCTION

E-commerce sites have gained prominence. Entrepreneurs launch firms on multiple platforms. Most organizations, enterprises, and government agencies use it to increase global trade efficiency. It dominates world business. During the epidemic, e-commerce boomed. Online transactions are popular. Card payments are widespread currently. Online credit card transactions are a big cause for e-development. commerce's Top corporations and institutions utilize credit cards to save time and streamline transactions [1]. This paper uses machine learning to eliminate fraud. Automated fraud detection systems identify, terminate, and forecast fraud. Digital payment systems are now making it simpler to receive payments. In the digital payment system, there is some unauthorized activity [2] done during the payment. It might be fraud, stealing the payment data, or some other intention. There is a group of people trying to scam the money. It involves stealing the money without knowing the person. It is a threat to the finance sector, business interaction, and business growth, and it is an illegal activity to gain financial benefit without knowing anyone in the act of spyware [3]. We are not only attempting to avoid fraud situations in this research, but we are also attempting to detect fraudulent conduct using ML approaches. Credit card fraud can occur not just at payment times, but also through cloning the cards. Cloning the cards resulted in around 25 lakh takas (BDT) being stolen from the card holder's account in the most recent year [3].

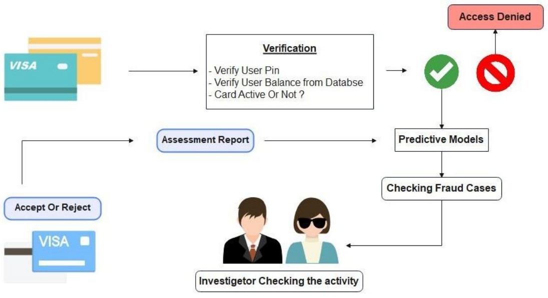

Fig. 1. Process Activity

Humans can't identify fraud. Credit card theft and loss are common types of fraud. This study utilizes machine learning to predict fraud in credit card transaction data sets. To resolve the imbalance between classes, logistic regression, random forest, and XGBoost will be used to the data set and the outcomes described.

II. LITERATURE REVIEW

The author examined the training and testing using Weka (Waikato Environment for Knowledge Analysis). To train and evaluate the models, the author utilized 0, 5, 10, 15, 20-fold cross validation. They employed decision tree, K-nearest neighbor, neural network, and logistic regression. Logistic regression was the best model [5].

The author examined an experimental study with imbalance classification problems. The experiment was done by using 8 machine learning classification algorithms. According to the study, SVM and ANN are the most preferable for imbalance classification approaches. Using SVM and ANN the results are Accuracy is 96% and 96%, Sensitivity is 39% and 47% respectively. Compared the hybrid of KNN and Naive Bayes classifiers with other mechanisms [6].

The author found out when it comes to detecting credit card deceptions, the hybrid classification model outperforms the voting-based classification technique, the accuracy is around 90%. Moreover, the execution time was the lowest for the hybrid classification [7].

The author's approach for enhancing classification accuracy is based on the concept of user split in which they classified users into old and new individuals before applying Cat Boost and Deep Neural Network to each category. The result that was obtained by their model: AUC (Area under the curve) score was 0.97 for Cat Boost and 0.84 for Deep Neural Network respectively [8].

The author analyzed credit card fraud data using naive Bayes, k-nearest neighbor, and logistic regression. Under and over sampled unbalanced datasets. Analyzing model's accuracy, sensitivity, specificity, precision, Matthews' correlation coefficient, and balanced classification rate determined the outcome. Nave Bayes, k-nearest neighbor, and logistic regression classifiers have optimum accuracy of 97.92%, 97.69%, and 54.86% [9].

The author used three models (54.27 % K-means, 99.70 % isolation forest, and 99.88 % logistic regression). Both K-means and neural networks were surpassed by logistic regression. It combines Bayesian hyper parameter optimizations to match the LightGBM parameters [10].

The author proposed system produces 98.40% accuracy, 92.88% areaunder receiver operating characteristic curve (AUC), Precision of 97.34%, and 56.95% F1-score. The most suitable algorithm that accurately detects fraud or outliers using supervised and unsupervised machine learning algorithms [11].

The author provided a comprehensive review of techniques for effectively detecting credit card fraud in both online and offline transactions. With the increasing use of credit cards for cashless transactions, the risks of credit card fraud have also grown. The paper emphasized the need for robust fraud detection methods and explores the use of data mining techniques and algorithms, specifically Hidden Markov Models (HMM) and Neural Networks (NN), to identify fraudulent activities. The working process likely involved data collection, preprocessing, feature extraction, model training, model evaluation, and fraud detection. However, specific details are not provided. The limitations of credit card fraud detection techniques, such as evolving fraud techniques, imbalanced datasets, false positives and negatives, and computational complexity, were also discussed [12].

The literature review highlights common limitations in credit card fraud detection, including imbalanced datasets, trade-offs between sensitivity and specificity, evolving fraud techniques, computational complexity, and data preprocessing issues. Limited discussion on external validation and generalizability is also observed. To address these limitations, we are employing various ML models for comparative analysis. This helps balance accuracy and complexity, adapt to evolving fraud techniques, and enhance interpretability. SMOTE and ADASYN classifications is used to tackle imbalanced datasets by generating synthetic samples of the minority class. These techniques aim to improve fraud detection performance, especially for underrepresented fraudulent transactions.

III. METHODOLOGY

As we mentioned earlier, Machine learning is a computer algorithm that can operate autonomously utilizing vast quantities of data. The application of machine learning for data analysis and control has grown in recent years. Supervised machine learning models are utilized to identify credit card fraud. Several data categorization methods, including oversampling, under sampling, power transformation, stratified, and repeated kfold cross validation, are used in this paper. During analyzing and clustering data from a dataset, unsupervised learning approaches are employed to do so. Other algorithms including neural networks, k-means clustering, probabilistic clustering approaches, and more are used in unsupervised learning.

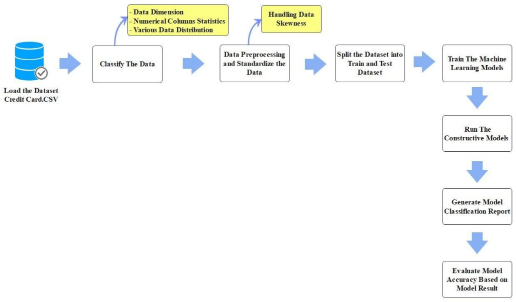
- Data Dimension
- Numerical Columns Statistics
- Various Data Distribution
Handling Data Skewness
Load the Dataset Credit Card.CSV
Classify The Data
Data Preprocessing and Standardize the Data
Split the Dataset into Train and Test Dataset
Train The Machine Learning Models
Run The Constructive Models
Generate Model Classification Report
Evaluate Model Accuracy Based on Model Result

Fig. 2. Methodology Workflow

- **Data Collection:** Analytical data collecting techniques are often utilized on a massive scale of data capture and processing. We are exploring methods for identifying credit card fraud in this study. In this circumstance, we must collect data that provides us with relevant information related to credit card transaction data. When it comes to financial concerns, such as unauthorized credit card usage, the situation becomes more complicated. The dataset used in this work was created for a research effort by World line and the Machine Learning Group of the University libre de Bruxelles. Furthermore, the dataset was made available on Kaggle, a community of data scientists and machine learners. This data collection contains September 2013 credit card purchases made by European cardholders. This dataset contains 492 fraudulent transactions out of a total of 284,807 transactions that occurred over the period of two days. Positive transactions account for just 0.17 percent of all transactions, indicating a huge imbalance in the dataset.

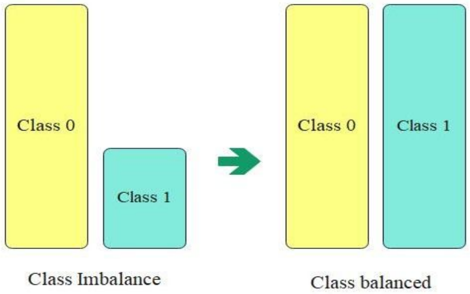

Fig. 3. Class imbalance

- **Data Classification:** Data classification is the process of putting data into categories that make it easy to find, sort, and store for future use. To make categorizing easier, we put observations into groups based on the data that was used to train the system. According to the data sets, a transaction with a value of 0 is real or legal, while a transaction with a value of 1 is fraudulent.
- **Class Imbalance Problem:** Using credit card fraud detection, it is feasible to identify the class imbalance problem that classifiers confront when presented with datasets in which the number of negative examples greatly out numbers the number of positive instances.
- **Re-sampling Approach:** As previously described in the section on the class imbalance issue, we are using re sampling methods, such as under sampling, oversampling, power transformation, SMOTE, and ADASYN, to fix the class imbalance problem. Resampling is a technique for retrieving multiple samples from the same data set. This is a non-parametric statistical inference approach.
  1. Under sampling is the process of randomly removing samples from the majority class from the training dataset. Most of the class instances are randomly dropped until a more balanced distribution is achieved with the random under sampling method. Random under sampling is the opposite of random oversampling. This method selects and eliminates samples from the majority class at random, reducing the number of examples of majority class data that are updated. Large volumes of data can be removed through random under sampling.
  2. Oversampling is the process of randomly adding duplicate cases from a minority class to a training dataset. Random oversampling chooses a random example from a minority class and replaces it with numerous copies of that instance in the training data, allowing a single instance to be chosen many times.
  3. To make the distribution more Gaussian, the Power Transformer package is included in the prepossessing library offered by Sklearn. This is often referred to as removing a skew from the distribution, although it is more accurately stated as stabilizing the variance of the distribution in most cases.
- **Synthetic Minority Over-sampling Technique (SMOTE):** The proposal of an over-sampling methodology that over samples the minority class by using synthetic samples rather than oversampling. This approach aims to create new minority classes by interleaving numerous minority cases in the space between feature vectors, rather than by using oversampling by replacement. The total number of artificial observations is therefore limited by the amount of data points sought by the SMOTE algorithm. This paper describes the correct operation of the SMOTE algorithm. Further, the problem of over-fit training data is solved using the SMOTE approach.

1. Calculate 2 feature vector points and multiply them by a random value between 0 and 1. Place a new data point along the dashed line separating two feature vectors.
2. Continue the method, and the number of synthetic observations will increase according to the number of SMOTE- created data points.

The hypothesis argues our financial data set is highly skewed. We use SMOTE techniques to improve accuracy throughout this case. As an example of the SMOTE methods in action, we may utilize a scatter plot to compare the initial and final data classifications.

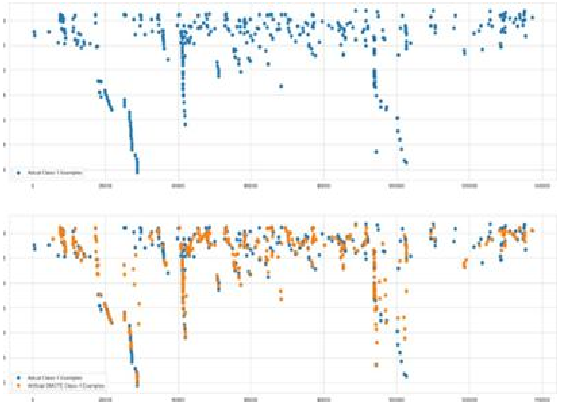

Fig. 4. SMOTE Oversampling

- **Adaptive synthetic**: These algorithms are often used when working with imbalance datasets. Adaptive synthetic (ADASYN) sampling strategy for learning from unbalanced data sets was described in prior research. ADASYN's core concept is to utilize a weighted distribution for distinct minority class examples according to their degree of learning difficulty, with more synthetic data created for minority instances that are more difficult to learn than those that are simpler to learn. As a result, this strategy's primary purpose is to avoid the repetition of any minority discovered in the used dataset. By using ADASYN to generate synthetic samples for the minority class in an adaptive manner, it is feasible to reduce the bias caused by an unequal data distribution.

In the context of an approach based on machine learning, logistic regression is an example of a supervised learning model. Calculations are made to determine the likelihood of a binary output of either 0 or 1.

- **Logistic Regression:** In the context of credit card fraud detection, a value of 0 indicates a legitimate transaction. Most of the time, logistic regression is used to data to make predictions about the binary output, which reveals whether such a transaction was false.
- **K-Nearest Neighbors (KNN)**: A simple Euclidean distance calculation is all that is required. KNN is utilized to tackle classification problems in the KNN model construction, which is primarily used to identify the data points that will be in the proper place.
- **Tree Model:** The tree model with Gini criterion is mostly used for data categorization and regression problems. It is also known as the Decision Tree model. It is a real-world application of supervised learning. In the tree model, the labels for each class are represented by the tree branches or tree leaves. Subtracting the total of the square probabilities of each class generates the Gini Index. It can demonstrate in a coherent way that the Gini index determines the chance of a randomly picked class being categorized arbitrarily.
- **Random Forest:** The first thing that is done in a random forest is to choose n records at random from a data collection that includes k records. This is done so that the results of the random forest are more accurate.
- **XGBoost:** It was built with the goal of maximizing efficiency. It does this by putting machine learning algorithms into action within the context of the Gradient Boosting framework. XGBoost is an implementation of parallel tree boosting, which is also known as GBDT or GBM, and it is used to address a broad variety of data science issues in a timely and accurate manner. The method works well on both balanced and unbalanced datasets.
- **Support Vector Machines (SVM):** The tasks of classification, regression, and finding outliers are all handled by support vector machines, which are supervised learning algorithms. SVMs are distinct from conventional classification methods in the sense that they determine the decision boundary in a manner that increases the distance between the closest data points for all classes. This contrasts with conventional classification methods, which determine the decision boundary in a manner that maximizes the distance between the closest data points for just one class.

IV. MODEL PERFORMANCE

The main task of this whole study is to detect the fraudulent transitions, for this, the accuracy and recall scores number should be very high. Moreover, the ROC value should be high as well.

- **Model 1:** Logistic regression using L1 and L2, test dataset accuracy levels were 0.9989 and 0.9990, which are high and may be considered an excellent result for fraudulent. The recall score ranges from 0.7385-0.7681. With these values we can say that Logistic regression L2 will be much preferable.
- **Model 2:** Tree model with Gini criteria and Tree model with Entropy criteria, if we look at the results accuracy level is same which is 0.9990 and recall scores are little different which are 0.6931 and 0.6939. ROC value with Tree model with Gini criteria is 0.8643 and Tree model with entropy criteria is 0.8539. With these values it is hard to say with is better than the other as there is just a slight difference in all the scores.
- **Model 3:** For KNN, the level of accuracy is 0.9993, which is pretty good, but the recall score for test

datasets is 0.8434, which is somewhat low, and the ROC value is 0.9008.

- **Model 4:** For XGBoost, accuracy level is 0.9995 which is high and recall score for test datasets is 0.9114 which is quite high and Roc value is 0.9759.
- **Model 5:** For SVM, accuracy level is 0.9986 which is quite high and recall score for test datasets is 0.5743 which is very low and Roc value is 0.8919. If we compare Model 3, Model 4 and Model 5, it can be clearly said that Model 4, XGBoost, will have the most preferable votes as the other models scores are low in many cases and cannot be trusted fully to detect the fraudulent. On the other hand, undoubtedly, Model 4 will be chosen for the excellent high scores.

V. MODEL OUTCOME

Table 1 below demonstrates that random forest and XGBoost performed better than other models in repeated KFold Cross Validation based on their total f1 score. This demonstrates how effective ensemble approaches are and how they may enhance performance despite a class imbalance. When random Under sampling was used, every model had a high recall score but a horrible accuracy score. SMOTE improved the accuracy scores for random forest and XGBoost, but the recall scores decreased significantly. When the identical re sampling procedure was utilized, however, XGBoost passed to increase the accuracy score. In conjunction with random forest, Repeated KFold Cross Validation yielded high accuracy scores of 0.96 and recall scores of 0.80, as well as an area under the ROC curve of 0.94. The 0.87 f1 score was also excellent. Random forest in repeated KFold Cross Validation were used so that we could decide.

From all the results it has been seen XGBoost and Random Forest scored better than every other technique. XGBoost was passed to improve the accuracy score. Repeated KFold Cross Validation, in combination with random forest, gave good accuracy and recall scores of 0.96 and 0.80, as well as a 0.94 area under the ROC curve. The f1 score of 0.87 was likewise great.

TABLE I. RESULT SUMMARY

Class	**Method's**	**P**	**Recall**	**f1-score**	**ROC**
LR(L2)	ADASYN	0.28	0.85	0.42	0.98
	SMOTE	0.30	0.85	0.44	0.98
	Random U. S	0.10	0.84	0.19	0.97
	Random O. S	0.23	0.86	0.37	0.98
	Power Transformer	0.77	0.55	0.64	0.97
	Stratified Kfold	0.23	0.86	0.37	0.98
	Repeated Kfold	0.93	0.57	0.70	0.98
KNN	ADASYN	0.0	0.0	0.0	0.50
	SMOTE	0.0	0.0	0.0	0.50
	Random U. S	0.0	0.0	0.0	0.50
	Random O. S	0.0	0.0	0.0	0.50
	Power Transformer	0.74	0.5	0.56	0.97
	Stratified Kfold	1.0	0.16	0.28	0.96
	Repeated Kfold	0.8	0.04	0.08	0.64
DT(Gini)	ADASYN	0.03	0.01	0.02	0.50
	SMOTE	0.08	0.02	0.03	0.51
	Random U. S	0.02	0.82	0.05	0.88
	Random O. S	0.0	0.0	0.0	0.50
	Power Transformer	0.69	0.73	0.71	0.85
	Stratified Kfold	0.67	0.65	0.66	0.83
	Repeated Kfold	0.71	0.74	0.73	0.87
DT(Entropy)	ADASYN	0.73	0.70	0.72	0.85
	SMOTE	0.69	0.70	0.70	0.85
	Random U. S	0.03	0.86	0.05	0.90
	Random O. S	0.81	0.69	0.75	0.85
	Power Transformer	0.69	0.71	0.70	0.85
	Stratified Kfold	0.8	0.65	0.72	0.83
	Repeated Kfold	0.72	0.78	0.75	0.89
RF	ADASYN	0.98	0.63	0.77	0.95
	SMOTE	0.96	0.65	0.78	0.94
	Random U. S	0.16	0.81	0.26	0.98
	Random O. S	1.00	0.63	0.78	0.94
	Power Transformer	0.91	0.71	0.80	0.93
	Stratified Kfold	0.99	0.67	0.80	0.94
	Repeated Kfold	0.96	0.80	0.87	0.94
XGBoost	ADASYN	0.94	0.77	0.84	0.97
	SMOTE	0.95	0.77	0.85	0.97
	Random U. S	0.11	0.85	0.20	0.98
	Random O. S	1.00	0.70	0.83	0.98
	Power Transformer	0.91	0.75	0.82	0.98
	Stratified Kfold	0.97	0.69	0.81	0.98
	Repeated Kfold	0.95	0.77	0.85	0.98
SVM	Random U. S	0.002	0.30	0.003	0.50
	Power Transformer	0.57	0.60	0.59	0.89
	Stratified Kfold	0.001	0.72	0.003	0.69
	Repeated Kfold	0.0	0.0	0.0	0.55

In the analysis, we employed a random forest and conducted KFold Cross Validation. The techniques we used for balancing the dataset are: Random Oversampling with StratifiedKFold CV, SMOTE Oversampling with Stratified KFold CV, and ADASYN Oversampling with Stratified- KFold CV, Random Under-sampling, Repeated KFold Cross Validation, StratifiedKFold Cross Validation, Repeated KFold Cross Validation and Power Transformer. Moreover, to tackle the class imbalance problem we implemented the boosting algorithm approach. For boosting we used the XGBoost algorithm. Tree model with entropy and Gini criteria was also used for the nodes having multiple classes. Besides all these models, Logistic Regression was used as well to compare with other models to get the best possible results at the end.

This study proposes an innovative and complete solution to decision-making class imbalance. The research compares Random, SMOTE, and ADASYN oversampling approaches with various cross-validation methods. The research uses XGBoost for ensemble learning, improving predicted performance. Decision tree models using entropy and Gini criteria demonstrate careful model selection. Logistic Regression benchmarking provides a unique perspective. Power Transformation preprocessing and a strong Random Forest decision-making framework augment the research. Evaluation criteria and real-world applicability are carefully considered, giving the article a useful guide for addressing class imbalance and making educated judgments in actual applications.

VI. CONCLUSION

This research paper explores the domain of credit card fraud detection, providing a comprehensive analysis of a system specifically designed to differentiate between legitimate and fraudulent transactions. The basis for this study is a dataset obtained from the public domain of Kaggle, which includes a significant number of 284,807 credit card transactions including monetary transfers. The empirical findings demonstrate that the XGBoost and Random Forest algorithms exhibit advantages over other strategies. To enhance precision, the integration of XGBoost was carefully implemented. By using the effectiveness of Repeated K-fold Cross Validation in conjunction with the Random Forest model, notable results were achieved. The accuracy and recall scores reached 0.96 and 0.80, respectively, demonstrating a high level of performance. Additionally, a significant area under the ROC curve was obtained, measuring at 0.94. The robustness of the F1 score, which was measured at 0.87, provides further evidence supporting the effectiveness of this strategy.

The primary focus of this investigation, however, is the examination of the significant disparity in social class that is inherent within the selected dataset. The present study addresses this challenge by employing a range of balancing techniques, such as Random Oversampling with stratified K-fold Cross Validation, SMOTE Oversampling with stratified K-fold Cross Validation, ADASYN Oversampling with stratified K-fold Cross Validation, Random under sampling, and multiple iterations of Repeated Fold Cross Validation. These efforts are in line with the essential need for achieving balance in datasets, which is necessary to guarantee the validity of the analysis and the accuracy of the findings made.

However, these advancements, the study admits a limitation that has quietly influenced its course. One noteworthy constraint of the study was the use of a dataset obtained from Kaggle, as opposed to using an internally created dataset. Although this hindrance temporarily hindered advancement, it eventually did not prevent the achievement of respectable results.

REFERENCES

[1] P. T. Praj Save, A Novel Idea for Credit Card Fraud Detection Using decision Tree, International Journal of Computer Application, p. 161,2017.

[2] H. T. M. A. a. A. B. Thanh Thi Nguyen1, Deep Learning Methods for Credit Card Fraud Detection, IEEE, India, 2020.

[3] S. Rahman, Credit, Debit Cards: Swindling on the rise, Fraudsters clone cards with data skimmed from shops, p. 1, 13 August 2017.

[4] W. Jolly, Common credit card frauds and how to avoid them, p. 1, 10 July 2019.

[5] B. G. G. V. R. S Venkata Suryanarayana, Machine Learning Approaches for Credit Card Fraud Detection, May 2018.

[6] S. M. Zainab Assaghir, An Experimental Study With Imbalanced Classification Approaches for Credit Card Fraud Detection, Vols. IEEE Access PP (99):1-1, July 2019.

[7] S. k. Darshan Kaur, Machine-Learning Approach for Credit Card Fraud Detection (KNN Naive Bayes), p. 5, 30 March 2020.

[8] T. D. C.-H. N. T. T. Nghia Nguyen, A Proposed Model for Card Fraud Detection Based on Cat Boost And Deep Neural Network, 19 April 2022.

[9] G. S. Vaishnavi Nath Dornadulaa*, Credit Card Fraud Detection using Machine Learning Algorithms, International conference on recent trends in advanced computing, 2019.

[10] J. O. Awoyemi, A. O. Adetunmbi and S. A. Oluwadare, Credit card fraud detection using machine learning techniques: A comparative analysis, 2017.

[11] n. D. Q. A. Kaneez zainab, A novel technique to Defraud credit card using an optimized cat boost Algorithm, vol. 100.4, no. 28th February 2022.

[12] R. RAJAMANI and M. RATHIKA, Credit card fraud detection using hidden Markov model, ijana.in, https://ijana.in/Special%20Issue/file38.pdf.

[13] Kumar, Ashish & Soni, Shivank & Agrawal, Chetan, A Survey Paper On Credit Card Fraud Detection Using Different Classifiers. International Journal of Computer Sciences and Engineering. 7. 552-559. 10.26438/ijcse/v7i1.552559, 2019.