International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019

Credit Card Fraud Detection using Machine Learning

P. Sai Gowtham Kumar, P. A. Sumanth Reddy, A. Mary Posonia

Abstract: ***Fraudulent transactions using credit card has been a growing concern with far reaching among various such as including government, corporate organizations, finance industry. Internet business is the most helpful answer for grow the client base and accomplish the biggest stage with a little venture. The fast development in the E-Commerce has significantly expanded Visas use for online buys and it actuated blow-up in the Credit card misrepresentation. For both online just as ordinary buy Credit card turned into the most well-known method of instalment, extortion cases associated with it are additionally emerging. The false exchanges are mistaken for certified exchanges and the basic example coordinating methods are not frequently enough to identify those cheats precisely. Effective location misrepresentation framework execution wound up basic to limit their misfortunes for all credit card issuing banks. Present day strategies dependent on Artificial Intelligence, Data mining, Fuzzy rationale, Machine learning, Sequence Alignment, Genetic Programming and so forth., are developed in distinguishing different Visa deceitful exchanges. When credit card transactions become a common mode of payment, machine learning has been based on handling the credit card fraud problem. This paper investigates naïve bayesian, k-nearest neighbor's performance on highly skewed credit card fraud based on genetic and optimization algorithm to determine the fraudulent transaction using credit card. Logistic Regression is a supervised classification technique which returns the probability of binary dependent variable predicted from the independent dataset variable that is logistic regression predicts the probability of different outcomes that have two values either yes or no and false or true. The Proposed System have been applied with genetic and optimization algorithm to find out the fraudulent transaction using credit card.***

Keywords: Genetic & Optimization Algorithm, Regression, Machine Learning

I. INTRODUCTION

In 2017, Credit and prepaid cards created over $31 trillion in absolute volume worldwide with misrepresentation misfortunes coming to over $21 billion. In that equivalent year there have been quite 225 billion get exchanges, assume that is anticipated to outperform 600 billion by 2025. Today, most Fraud Detection Systems (FDS) keep on utilizing progressively complex machine learning calculations to learn and distinguish false examples continuously, just as disconnected, with insignificant unsettling influence of certified exchanges. Today, many Fraud Detection Systems (FDS) tend to use increasingly modern machine learning calculations to constantly, just as disconnected, learn and recognise fake examples with negligible unsettling impact of true exchanges. For the most part, FDS needs to address some intrinsic difficulties found with the undertaking: outrageous data set unevenness as fakes talk only to a compact amount of utter transactions, conveyances that grow due to changing buyer actions and challenges of assessment that surround on-going information handling. In credit card transactions, we designed a system to detect fraud. The system provides most of the essential features required to identify fraudulent and legitimate transactions. When technology changes, tracking the behavior and pattern of fraudulent transactions becomes difficult. We just noticed the fraudulent activity, but we didn't stop it. It is not easy in real time to prevent known and unknown fraud, but it is feasible. The proposed architecture is basically designed to detect online payment credit card fraud, and emphasis is placed on providing a system of fraud prevention to verify a transaction as fraudulent or legitimate. It is assumed that the issuer and the acquirer bank are linked with each other for implementation purposes. To enforce this program in real-time scenario, sharing best practices and increasing consumer awareness among people can be very helpful in reducing the losses caused by fraudulent transactions.

Revised Manuscript Received on December 30, 2019.
* Correspondence Author
P. Sai Gowtham Kumar, Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India
P. A. Sumanth Reddy, Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India
A. Mary Posonia, Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

II. RELATED WORKS

Dwindle Martey Addo [1], proposed a model on true informational index in recognizing the misrepresentation utilizing machine learning. The essential ten things are accumulated and they are utilized in testing the dependability and execution in various information. They found that tree based model is increasingly proficient and stable. There emerges numerous inquiries identifying with profound learning. Profound learning is adding of layers to neural systems regardless of whether they are imitated. Generally these profound learning rely upon 4 sorts of models: Neural systems utilizing loads, Neural systems in same time, neural systems without time factor and standard neural systems which is blend of different kinds with no structure.

Fahimeh Globoid and Mohsen Ghobadi [2] proposed a framework which uses neural system for the cost-touchy demonstrating of extortion information. The proposed framework can control the loss of information by utilizing Artificial Neural Networks.

It likewise consolidates meta-cost strategy which can lessen the hazard related with the method. Neural systems are being used for misrepresentation recognition from recent years. There can be both extortion and unique information in the experiment which is taken for analysis. It turns into a testing issue to recognize these two. The framework they have proposed can make sure of this issue. The framework likewise gives entirely great recognition rate contrasted with different frameworks. The principle downside of the framework was it was not able handle oddity recognition approach. Additionally, the adequacy of the framework can be enhanced with the utilization of Meta heuristic methodology.

Snehal Patil [3] et al. they utilized expandable systems to distinguish the information that are utilized in exchange to ascertain the extortion in less measure of time which is a key issue in the field of online business. Aside from expandability and effectiveness, the extortion recognizing work has some specialized inconvenience that incorporates slanted preparing information and cost per mistake which isn't uniform. In recognizing misrepresentation machine learning assumes

III. PROPOSED SYSTEM

First the credit card dataset is taken from the supply, and cleaning and approval is executed on the dataset which joins disposal of excess, filling void territories in sections, changing imperative variable into components or exercises then actualities is part into 2 sections, one is preparing dataset and another is check data set. Presently k crease move approval is done that is the special example is arbitrarily divided into k same and equivalent measured subsamples. Of the k subsamples an individual subsample is held in light of the fact that the approval actualities for looking at the model and the last k-1 subsamples are utilized as instruction data, designs are made for Logistic relapse, choice tree, SVM, Random woodland after which exactness, affectability, explicitness, accuracy are determined and a differentiation is made. The dataset is sourced from ULB framework contemplating association. The dataset incorporates Visa exchanges and event of exchanges that occurred in days is exhibited on this dataset, comprehensive of 284,786 exchanges. The dataset is outstandingly unequal and skewed towards the magnificent tastefulness and extortion cases make up 0.173% of the exchanges certainties. It conveys handiest numerical (constant) enter factors that are a result of a Principal Component Analysis (PCA) trademark decision change coming about to twenty-eight basic added substances. Furthermore, generally speaking of 30 input highlights are connected. Social element of the card is demonstrated by method for a variable of each profile usage speaking to the spending behavior of the clients alongside days of the month, hours of the day, topographical areas, and diverse kinds of clients. where the exchange happens in the area. Later those factors are used in making renditions which separate deceitful exercises. The subtleties and the verifiable past certainties about the capacity are not issued in view of security issues. The time factor monitors the quantity of seconds that has been conveyed in each exchange alongside exchange that is done toward the start in the dataset. The overall component diagram is displayed in the figure 1.

IV. SYSTEM ARCHITECTURE

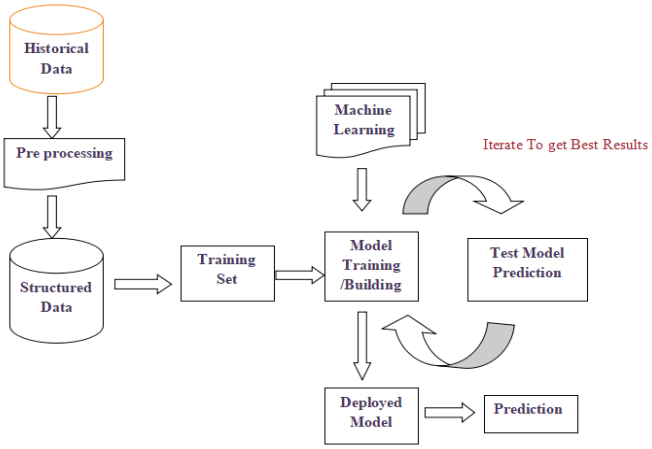

Fig. 1 Proposed system Architecture

V. METHODS OF IMPLEMENTATION

In Existing System Fraud Detection, the massive stream of payment requests is quickly scanned by automatic tools that confirm that transactions to authorize. This system Provide many machine learning algorithms that has used for fraud detection depend upon expectations that hold in an exceedingly world fraud detection system. This system made lack of realism issues two main aspects like the time and way with that supervised info is provided and therefore the measures wont to assess fraud-detection performance is non consistent. Effective learning approach to address the issues, including the latency of validation and the communication of feedback. In a large number of credit card transactions, this learning strategy is tested. The proposed architecture is essentially designed to detect online payment credit card fraud, and emphasis is placed on providing a system of fraud prevention to verify a transaction as fraudulent or legitimate. For implementation purposes it is assumed that issuer and acquirer bank is connected to each other. This paper analysis the execution of naïve bayesian, knn and logistic regression on highly skewed credit card fraud based on genetic and optimization algorithm to find out the fraudulent transaction using credit card .Dataset of transactions using credit card is sourced. A hybrid technique of under-sampling and oversampling is disbursed on the inclined knowledge the three techniques are applied on the raw and preprocessed data. The three techniques are applied on the raw and preprocessed data. The execution of the various techniques is evaluated based on sensitivity, accuracy, precision, specificity, balanced classification rate and Matthews correlation coefficient. The comparative results exhibit that k nearest neighbour performs higher than naïve bayesian and logical regression techniques we tend to Describe the mechanisms control a true world fraud detection technique FDS and supply a proper model of the effective classification downside to be self-addressed in fraud detection.

The performance measures that are considered in experimental. With realistic model, we propose an effective learning strategy for addressing the above challenges, including the verification latency and the alert feedback interaction.

1) Outline Data Analysis and Machine Learning Using Optimization Algorithm

Optimization Algorithms are organized through their interactions with the analysis of machine learning and knowledge. New algorithms and new interest in current algorithms; difficult formulations and new paradigms; revived stress on positive topics: optimization algorithms, complexity, structured and several new (excellent) machine learning researchers. Use this information to predict different similar knowledge. Extremely multidisciplinary space foundations in statistics, AI machine learning, parallel systems, database development provide a modeling / formulation toolkit and recursive techniques modeling and domain-specific information is very important "80% of knowledge processing is spent on enhancement methods and data preparation. The execution of the techniques is evaluated based on accuracy, sensitivity, specificity, precision. The results indicate about the optimal accuracy for logistic regression, decision. Even fraudulent transactions found by investigators that show patterns involving few elements of the feature vectors will be found based on training sets. The performance measures should also take into account the capacity of the investigators, as they must review all the alerts raised by the fraud detection system

2) Matrix Completion of Data Analysis

Noise or errors in aj and yj. would like φ (and x). We wish to generalize solutions to the discovered info that may be achieved by regularized formulations is shown within the figure 2.

Avoid overfitting: Certain knowledge is seen as an associated empirical degree, a sampled illustration of some underlying reality. Would like to avoid the actual sample being overfitted. (Training needs to produce the same outcome for different samples from the same information set) again, generalization / regularization.

Missing data: Vectors aj is also missing a number of its components (but contains useful data)

Missing labels: There is also a loss of some or all yj or null semi-supervised or unsupervised learning.

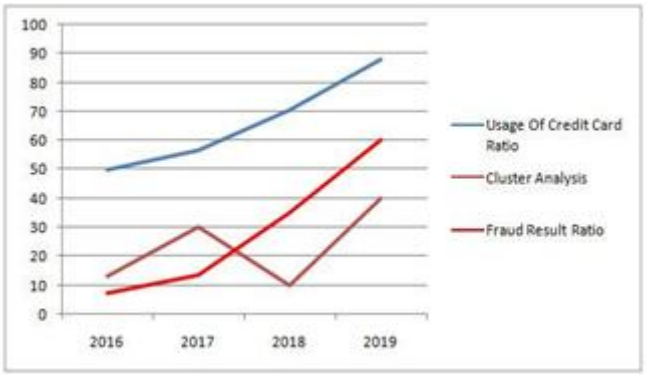

Fig. 2 Usage of Credit card Analysis

VI. ANALYSIS AND RESEARCH

The importance of Machine Learning is improved over contribution Analysis. In this analysis train machines to be trained in a way to outline situations establish associated label events or predict a value within the present data or future data science Is of the essence it's essential to check the underlying data and model it by choosing an applicable algorithm to approach any such use case the varied management parameters of the algorithm must be tweaked to suit the data set. As a result, is shown on the figure 3, the developed application improves and becomes more efficient in solving the problem. we Have tried for instance the modeling of a data set employing a machine learning paradigm classification with credit card fraudulent transaction detection being the base Classification could be a machine learning paradigm that involves account a perform that may separate data classes or categories the obscurity of the net makes it a perfect venue for credit card thieves "a sizable amount of on-line purchases during a short amount of time is additionally seemingly to urge a credit card account flagged " he afore mentioned Multiple purchases in fast succession also will go away the credit card companies'

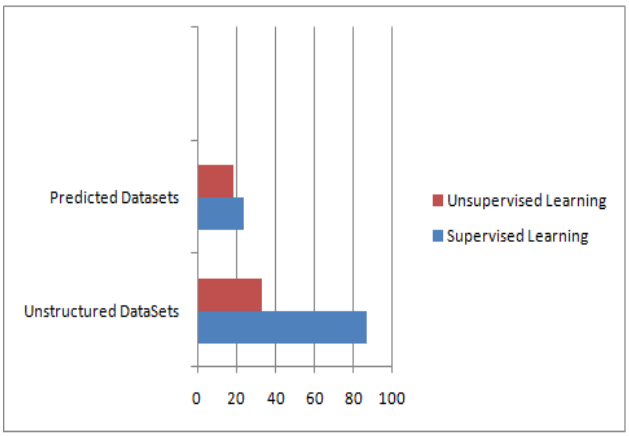

Fig. 3 Comparison of Supervised Vs Unsupervised Learning

VII. METHODOLIGES

1. Supervised learning and unsupervised learning

Using supervised approach helps to find out about the label of past transactions, they appear to not identify the pattern of fraud that occurred in past historical records. While unsupervised technique helps to determine the type of transaction

Least Square Method

min x f (x) := 1 2 Xm j=1 (a T j x − yj) 2 = 1 2 kAx − yk 2 2

Here the function mapping data to output is linear:

$$\varphi(aj) = a\ T\ j\ x.$$

The above methodize reduces sensitivity of the solution x to noise in y.

$$\min x\ 1\ 2\ kAx - yk\ 2\ 2 + \lambda kxk\ 2\ 2\ .$$

Thus the methodize generates solutions x with few nonzeros:

$$\min x\ 1\ 2\ kAx - yk\ 2\ 2 + \lambda kxk1.$$

Feature selection: Nonzero locations in x indicate important aj components. Nonconvex separable methods (SCAD, MCP) have nice statistical properties, but nonconvex set results.

Regression over a structured matrix: Observe a matrix X by probing it with linear operators

Aj(X), giving observations yj , j = 1, 2, . . . , m. Solve a regression problem:

min X 1 2m Xm j=1 (Aj(X) − yj) 2 = 1 2m kA(X) − yk 2 2 .

Each Aj may observe a single element of X, or a linear combination of elements. Can be represented as a matrix Aj , so that Aj(X) = hAj , Xi. Seek the "simplest" X that satisfies the observations. Nuclear-norm (sum-of-singular-values) methodize term that induces small rank on X: min X 1 2m kA(X) − yk 2 2 + λkXk, for some λ > 0.

2. Unstructured Data Analysis

Learning from an unbalanced dataset and using the sampling method for balancing it is quite challenging. A publicly accessible dataset that includes many cardholders transactions. The dataset includes minimal, highly imbalanced, fraudulent transactions. Therefore, under-sampling is shown in figure 4.

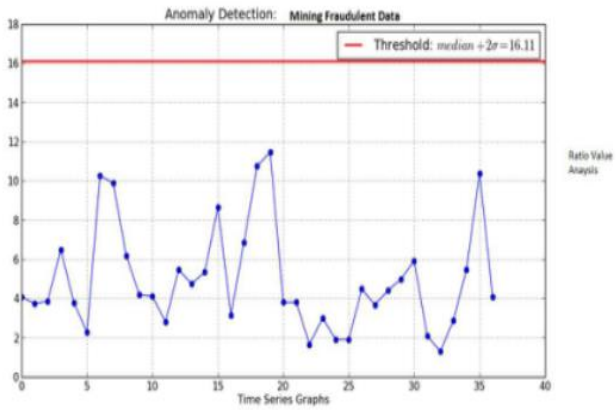

Fig. 4 Anomoly Detection

3. Fraud Detection Classifier

Data with theoretical and statistical characteristics can be handled by logistic regression. Decision Tree is a supervised method of learning that uses models extensively for tasks of regression and classification. Random Forest system used by decision tree collection for regression and classification.

VIII. DATASET DESCRIPTION

A dataset is an assortment of information. most typically a data set corresponds to the contents of one info table, or one applied mathematics knowledge matrix, wherever each column of the table represents a specific variable, and every row coinside with a given info set in question.

Table. 1 Sample Data set Table

	7	NODE_ID	CATEGORY	SUBCATEGORY	TYPE	FRAUDULENT_TRANS_NUM	TOTAL_TRANS_NUM	GROUP
75		Customer075	Customer	25to34	c	3	15	4
76		Customer076	Customer	25to34	c	1	16	4
77		Customer077	Customer	Under25	c	2	11	4
78		Customer078	Customer	45to54	c	4	17	4
79		Customer079	Customer	65andOver	c	2	9	4
80		Customer080	Customer	25to34	c	4	13	4
81		Merchant0001	Retail	DrugStores	m	0	1	1
82		Merchant0002	Retail	FoodStore	m	0	1	1
83		Merchant0003	Services	Restaurants	m	1	1	1
84		Merchant0004	Services	Restaurants	m	0	1	1
85		Merchant0005	Services	OtherServices	m	0	1	1
86		Merchant0006	Services	OtherServices	m	0	1	1
87		Merchant0007	Retail	General	m	1	1	1
88		Merchant0008	Services	OtherServices	m	0	1	1
89		Merchant0009	Retail	GasStation	m	1	1	1

IX. RESULT AND DISCUSSIONS

Transactions using credit cards benefits the merchant (seller) and the credit card provider. Credit card provider takes a fees which is close to 1% to 3% for each transaction generally (numbers are indicative, can vary across regions, type of merchants etc) which can be considered as profit in your case, considering you are modelling this for credit card providers. Let this fees be r. Also assume that all fraud transactions are realized by credit card provider only.

Expected Profit on a non-fraud transaction = Amount * r - (average cost per transaction, but lets ignore this as of now)

Expected Loss = E[L]=∑i=1nLiPi
=−0.05 (amount) 0.8+(amount) 0.2
=−.04 (amount)+(amount) 0.2=.16 (amount)

Expected loss for a fraud transaction = Amount Net expected profit for any transaction =P(non-fraud) * r * Amount - P(fraud) * Amount.

So ideally you should rank your transaction as per this equation. This will bring big amount transactions to top even if their probability of fraud is little lower, and will bring transactions with very high probability of fraud to top as well if transaction amount is not very low. Let us recognize if you wish to model for marchendiser (seller) , and I will try to answer for that as well. The two things you write are equivalent. Let pp be the probability of a fraudulent transaction, (1−p)(1−p)be the probability of a non-fraudulent transaction, and XX be the amount of the transaction.

qX−pX===:=(q−p) X((1−p)−p) X(1−2p) Xa XqX−pX= (q−p)X = ((1−p)−p) X=(1−2p)

X:=a X Note that (1−2p):=a(1−2p):=a, or we have defined aa to be (1−2p)(1−2p). Since pp is the only factor affecting aa, if you sort your list of pXpX from highest to lowest, you're going to get the same ranking as if you sorted by aXaX. (Well, the two lists will be exactly flipped, so if you sorted pXpX from highest to lowest, it's the same as sorting by aXaX from lowest to highest.) For the expected loss of a given transaction, it will be ∑ni=1LiPi∑i=1nLiPi, where LiLi is the loss of that transaction and PiPi is the probability of that instance occurring. Let's say a transaction is fraudulent 20% of the time. If Li=0Li=0 when the transaction is non-fraudulent but Li=amount Li=amount when that transaction is fraudulent (where amount is the amount involved in the transaction), then the expected loss will be:

Expected Loss=E[L]∑i=1nLiPi0 0.8+(amount) 0.2 (amount) 0.2.

In this case, your expected loss for a particular transaction would be 20% of that transaction, or 0.2(amount) 0.2(amount). However, if you actually gained a specific quantity of money if a transaction isn't fraudulent, then that must be considered. For example, if you gained 5% of the transactional amount on every non-fraudulent transaction, then your expected loss during this case would be, your expected loss would be 16% of that transaction, or 0.16(amount)0.16(amount). If you wanted to sum the expected loss over a set of transactions, you can calculate this for each transaction and sum the losses.

X. CONCLUSION

The result of the intended models in overall performance was superior. Overall results show that stacking classifier using Training Model, Meta classifier is most promising to predict fraud transaction in the dataset, followed by the classifier for data analysis. The paper focuses on this aspect of the Fraud Detection System and proposes a method for designing, learning, and upgrading the Datasets to boost the performance of fraud detection. Transactions are invariably related to feature vectors that have either received an outsized fraud score or a high chance of being a fraud generate alert. More number of alerted transactions are reported to the investigators, which represent the final layer of control. The assessment of the learning model is based on its accuracy recall, precision, specificity. The outcome of all the intended models in the overall performance was superior.

XI. FUTURE WORK

Further enhancement can be made by making this system secure by using both merchant and customer certificates and by adding new checks as technology changes can be added to know the pattern of fraudulent transactions and alert the respective card holders and bankers when fraud activity is identified. The dataset available on day to day processing may become outdated, it is compulsory to have updated data for effective fraud behavior identification. To this extent, the incremental approach is necessary in making the system to learn from past as well as present data and capable of handling the both. Fraudster uses different new techniques that are instantaneously growing along with new technology makes it difficult for detection. Also the nature of access pattern may vary from one geographical location to another (such as urban and rural areas) that may result in a false positive detection. In such a case a future enhancement may be based on new multiple models with varying access pattern needs attention to improve the effectiveness. Privacy preserving techniques applied in distributed environment resolves the security related issues preventing private data access.

REFERENCES

1. A. Dal Pozzolo, O. Caelen, Y. A. Le Borgne, S. Waterschoot, and G. Bontempi, “Learned lessons in credit card fraud detection from a practitioner perspective,” Expert Syst. Appl., vol. 41, no. 10, pp. 4915–4928, 2014.
2. M. A. Scholar, M. Ali, and P. Fellow, “Investigating the Performance of Smote for Class Imbalanced Learning: A Case Study of Credit Scoring Datasets,” vol. 13, no. 33, pp. 340–353, 2017.
3. H. He, W. Zhang, and S. Zhang, “A novel ensemble method for credit scoring: Adaption of different imbalance ratios,” Expert Syst. Appl., vol. 98, pp. 105–117, May 2018.
4. A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, “Calibrating Probability with Undersampling for Unbalanced Classification.”
5. Y. Sahin and E. Duman, “Detecting Credit Card Fraud by Decision Trees and Support Vector Machines,” Int. Multiconference Eng. Comput. Sci., vol. I, pp. 442–447, 2011.

6. V. Van Vlasselaer et al., “APATE: A novel approach for automated credit card transaction fraud detection using network based extensions,” Decis. Support Syst., vol. 75, pp. 38–48, 2015. [17] C. Phua, D. Alahakoon, and V. Lee, “Minority report in fraud detection,” ACM SIGKDD Explor. Newsl., vol. 6, no. 1, p. 50, 2004.
7. E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun, “The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature,” Decis. Support Syst., vol. 50, no. 3, pp. 559–569, 2011.
8. C. C. Lin, A. A. Chiu, S. Y. Huang, and D. C. Yen, “Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts’ judgments,” Knowledge-Based Syst., vol. 89, pp. 459–470, 2015.
9. N. Carneiro, G. Figueira, and M. Costa, “A data mining based system for credit-card fraud detection in e-tail,” Decis. Support Syst., vol. 95, pp. 91–101, 2017.
10. A. C. Bahnsen, D. Aouada, and B. Ottersten, “Example-dependent cost-sensitive logistic regression for credit scoring,” Proc. - 2014 13th Int. Conf. Mach. Learn. Appl. ICMLA 2014, pp. 263–269, 2014.
11. M. Carminati, R. Caron, F. Maggi, I. Epifani, and S. Zanero, “BankSealer: A decision support system for online banking fraud analysis and investigation,” Comput. Secur., vol. 53, pp. 175–186, 2015.
12. C. Whitrow, D. J. Hand, P. Juszczak, D. Weston, and N. M. Adams, “Transaction aggregation as a strategy for credit card fraud detection,” Data Min. Knowl. Discov., vol. 18, no. 1, pp. 30–55, Feb. 2009.
13. G. Rushin, C. Stancil, M. Sun, S. Adams, and P. Beling, “Horse race analysis in credit card fraud - Deep learning, logistic regression, and Gradient Boosted Tree,” 2017 Syst. Inf. Eng. Des. Symp. SIEDS 2017, pp. 117–121, 2017.
14. R. J. Bolton, D. J. Hand, F. Provost, L. Breiman, R. J. Bolton, and D. J. Hand, “Statistical Fraud Detection: A Review Comment Comment Rejoinder,” Stat. Sci., vol. 17, no. 3, pp. 235–255, 2002.
15. X.-Y. Liu, J. Wu, and Z.-H. Zhou, “Exploratory Undersampling for Class Imbalance Learning,” vol. 39, no. 2, 2009.
16. E. A. Mohammed, M. M. A. Mohamed, C. Naugler, and B. H. Far, “Toward leveraging big value from data: chronic lymphocytic leukemia cell classification,” Netw. Model. Anal. Heal. Informatics Bioinforma., vol. 6, no. 1, p. 6, Dec. 2017.
17. G. H. John and P. Langley, “Estimating Continuous Distributions in Bayesian Classifiers,” Feb. 2013.
18. S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, “Data mining for credit card fraud: A comparative study,” Decis. Support Syst., vol. 50, no. 3, pp. 602–613, 2011.

AUTHORS PROFILE

P. Sai Gowtham Kumar, Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

P. A. Sumanth Reddy, Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

A. Mary Posonia, Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India