JCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol. 10, Issue.12, December 2022 E-ISSN: 2347-2693

Detecting Fraudulent Transactions with the Ensemble Learning

Sayee Chauhan

Department of MultiDisciplinary Engineering, Vishwakarma Institute of Technology, Pune, India

Author's Mail Id: sayee.chauhan20@vit.edu

DOI: https://doi.org/10.26438/ijcse/v10i12.2327 | Available online at: www.ijcseonline.org

Received: 23/Oct/2022, Accepted: 26/Nov/2022, Published: 31/Dec/2022

***Abstract*—** Credit card companies must have the ability to identify fraudulent credit card transactions in order to stop customers from being charged for goods they did not purchase. These problems may be resolved with data science, and when combined with machine learning, it is extremely important. This study seeks to show how machine learning may be used to model a data set using credit card fraud detection. The Credit Card Fraud Detection Problem includes modelling prior credit card transactions using data from those that turned out to be fraudulent. Then, this model is used to analyse a new transaction to determine whether or not it is fraudulent. The objective is to detect 100% of the fraudulent transactions while minimising erroneous fraud categories. Due to the E-Commerce sector's recent explosive expansion, fraudulent credit card transactions have cost incredibly significant sums of money. An effective method to stop these fraudulent transactions is to use a strong model based on cutting-edge machine learning algorithms that can handle massive volumes of data while still producing precise findings. In this study, the effectiveness of decision trees, random forests, and linear regression for identifying credit card fraud is compared.

***Keywords*—** Outliers, Decision Tree, Confusion Matrix, Isolation Forest, Logistic Regression, Naive Bayes Classifier, Credit Card Fraud

I. INTRODUCTION

Fraud is defined by the Association of Certified as any intentional or planned act of robbing someone else of their money or property via trickery, deceit, or other unfair means.Add the complete name followed by the abbreviation CCF is a term used to describe the unauthorised CC procedure or information that has been deprived of its owner's data. Applications & behaviours of the CCF trick that differ from one another are linked to two types of fraud. Name the first and second groups specifically. When app fraud occurs, scammers use fraudulent or other information to ask for a new card from the bank or give it to businesses. A user may submit multiple applications using the same usual of descriptions (duplicate fraud) or another person may submit applications using the same normal of descriptions (named identity fraud).

One of the largest challenges facing commercial enterprises today is credit card fraud. There are two broad categories in which to place fraudulent transaction techniques. Theft of the credit card physically is one, while theft of data like the card number and CVV is another.

The automatic analysis of recorded transactions to look for fraudulent activity is one of the most widely investigated methods of fraud detection, according to Bolton et al. (2001). Transaction data, which includes a number of attributes, is kept in the service provider's databases whenever a credit card is used (such as the credit card identification, transaction date, recipient, and transaction amount). Rarely is the information from a single transaction enough to detect fraud, thus aggregate indicators like the total daily spending, the weekly transaction volume, or the average transaction value must be considered in the analysis. Algorithms built using fraud indications are typically used to detect fraud.

In credit card transactions, "fraud" refers to the unauthorised use of a credit card by a person who is not the account owner. The essential preventative measures can halt this misuse, and it is also possible to research the behaviour of such fraudulent operations to lessen recurrences and prepare for them. The use of another person's credit card for personal advantage when neither the cardholder nor the organisation in charge of issuing the card is known as credit card fraud, to put it another way. Monitoring user populations' behaviour is a crucial part of detecting fraud since it enables the identification, detection, and prevention of unwelcome behaviours including fraud, intrusion, and defaulting. Communities like machine learning and data science, where an automated solution is conceivable, should address this very important issue. This problem is particularly challenging from a learning perspective since it exhibits a number of traits, such as class imbalance. There are much more honest than dishonest trades.

Additionally, the transaction patterns' statistical properties frequently change over time. A fraud detection system's implementation in the real world is not without challenges, though.In real-world examples, the massive volume of payment requests is quickly analysed by automated tools to choose which transactions to approve. Machine learning algorithms are employed to analyse all authorised transactions and identify any that appear dubious. In order to determine whether the transaction was legitimate or fraudulent, investigators who are looking into these allegations get in touch with the cardholders. The automated system collects feedback from the investigators, which is then used to train and update the algorithm to progressively enhance fraud detection performance over time.

II. RELATED WORK

The definition of fraud is an unlawful or criminal deception intended to generate financial or personal gain. It is illegal to violate a law, rule, or policy with the goal to receive an unrecognised financial benefit. A lot of publicly available information has previously been published in this sector on the subject of anomaly or fraud detection. According to a comprehensive analysis conducted by Clifton Phua and his colleagues, some of the techniques employed in this sector include adversarial detection, automated fraud detection, and data mining applications. Suman, Research Scholar, GJUS&T at Hisar HCE, offered methods like supervised and unsupervised learning for credit card fraud detection in a different publication. Even though some of these techniques and algorithms achieved unexpected success, they were unable to offer a reliable, long-lasting answer to fraud detection. Wen-Fang YU and Na Wang presented a similar area of research in which they used distance sum algorithms, outlier mining, outlier detection mining, and outlier detection mining to precisely predict fraudulent transactions in an experiment simulating credit card transaction data from a particular commercial bank. Data mining's field of outlier mining is primarily utilised in the financial and internet sectors. It focuses on identifying detached objects from the main system, or transactions that aren't real. They have measured the difference between the observed value of an attribute and its preset value by using attributes related to consumer behaviour and basing their calculations on those attributes' values. Unusual methods like hybrid data mining/complex network classification algorithms, which are based on network reconstruction algorithm and allow creating representations of the deviation of one instance from a reference group, have typically proven effective on medium-sized online transactions. These methods are able to detect illegal instances in a real card transaction data set. Additionally, there have been initiatives to advance from a totally different perspective. The alert-feedback interaction in the event of a fraudulent transaction has been improved. A feedback would be delivered to the authorised system to deny the current transaction in the event of a fraudulent transaction. One method that provided new insight into this area dealt with fraud in a different way: Artificial Genetic Algorithm. It was effective in identifying fraudulent transactions and reducing the amount of false alarms. Even so, there was a categorization issue with fluctuating misclassification costs. The many sequential models and machine learning methods for fraud detection are reviewed in this section. Applications for several financial credit cards with transaction histories are reviewed.

Credit card transactions are significantly affected by the challenge of binary classification since they can either be classed as legitimate (true class) or valid (false class) (true class). Highly skewed information about credit card theft was examined by Awoyemi et al. in 2017 [1]. This research looks at the effectiveness of a number of methods, including Naive Bayes, KNN, and Logistic Regression. 284,807 credit card transaction-based data from customers in Europe were gathered. On the distorted data, a hybrid undersampling and oversampling technique is used. Python is used to perform three operations on the unprocessed and preprocessed data. The findings indicate that Naive Bayes, K-Nearest Neighbor, and Logistic Regression classifiers have the best accuracy, with respective values of 97.92%, 97.69%, and 54.86%. KNN outperforms Naive Bayes and Logistic Regression, according to the comparison results. Dal Pozzolo et al. proposed three significant additions in 2017 [2]. With the aid of their research and a formalisation of the fraud-identification problem, the authors first accurately represent the operational circumstances of FDSs, which regularly monitor enormous volumes of credit card transactions. The authors also gave illustrations of how to use the most effective evaluation metrics for fraud detection. Second, the authors created and tested a unique learning technique to handle class imbalance, idea drift, and verification latency. Third, to illustrate the impacts of class inequality and idea drift, the authors employed a real-world information stream with more than 75 million transactions authorised over three years. To educate the behavioural characteristics of typical and anomalous transactions, two different types of random forests are used. The approach proposed by Xuan et al. in 2018 [3] contrasted and analysed the effectiveness of numerous random forests with various classification models for detecting credit fraud. Data from these tests was provided by an online retailer in China. Long short-term memory networks were used by Jurgovsky et al. in their study from 2018 [4] to frame the fraud identification problem as a sequence classification challenge that also contained transactional sequences. Modern attribute aggregation techniques are also used by the system, and the framework's results are reported using traditional retrieval measures. The LSTM increases identification accuracy for offline transactions while the cardholder is physically present as compared to a benchmark Random Forest classifier. . The use of manual attribute aggregation techniques benefits both sequential and nonsequential learning systems. Following a review of true positives, it was discovered that both techniques had a tendency to spot various types of fraud, indicating that they should be used together. Varmedja et al. offered a variety of methods for categorising transactions as fraudulent or legal in their 2019 [5] study. The dataset utilised in the study was for identifying credit card fraud. The SMOTE method was employed to oversample the dataset because it was severely unbalanced. The dataset was divided into training and test halves and attributes were chosen. Logistic Regression, Random Forest, Naive Bayes, and Multilayer Perceptrons were the technologies employed in the study. The study demonstrates the accuracy with which each technology can detect credit card theft. Using the established framework, other anomalies might be discovered. The idea that patterns of fraud can be found by looking at previous transactions is the foundation of supervised learning-based systems for identifying credit card fraud. But because it must account for shifting customer behaviour and fraudsters' capacity to develop new fraud patterns, the process becomes complex. Unsupervised learning techniques can be used by fraud detection models to look for anomalies in this situation. Carcillo et al. in 2019 [6] proposed a hybrid methodology for improving fraud identification accuracy by combining supervised and unsupervised approaches. We analyse and evaluate unsupervised anomaly ratings generated on a real, labelled credit card fraud identification dataset using different granularities. The efficiency of the combination, which also improves identification precision, is supported by experimental findings. Machine learning approaches are employed in the study proposed by Randhawa et al. in 2018 [7] to identify credit card fraud. The application of conventional methods occurs first. The next step is to deploy hybrid tactics built on AdaBoost and public voting. A publicly accessible credit card dataset is utilised to assess the performance of the framework. The real-time credit card dataset of a financial institution is then used to evaluate the information. Additionally, distortion is added to the data samples in order to confirm the techniques' resilience. The outcomes of the experiment show that credit card theft may be reliably detected using the popular vote approach. De Sá et al. presented the Fraud-BNC approach in 2018 [8] to pinpoint credit card fraud issues. The proposed methodology is based on the Bayesian network classification model. Fraud-BNC was created impulsively using information from PagSeguro, the most popular online payment system in Brazil, and tested against two cost-sensitive classification methods. Seven other methodologies were compared to the obtained findings, and the methodology's cost-effectiveness and data classification problem were assessed. The most trustworthy methodology for striking a good compromise between the two points of view was Fraud-BNC, which increased the existing organization's financial efficiency by up to 72.64%. To detect fraudulent behaviour, Sailusha et al. created a model for identifying credit card theft in 2020 [9]. This study's main area of interest is machine learning. We employed both the Random Forest and AdaBoost techniques. The two methodologies' results are compared using their respective accuracy, precision, recall, and F1-scores. The confusion matrix is used to create the ROC curve. Performance metrics for these two techniques, such as accuracy, precision, recall, and F1-score, were compared. The performance metrics for a fraud detection methodology are thought to be the best. Economic fraud has made a name for itself as a threat and a significant component of the financial system. Data mining is a tactic that has been effective in identifying credit card fraud in internet transactions. It is challenging to detect credit card theft since the features of fraudulent and legitimate activity change over time and the datasets used are very biassed. Logistic Regression, Naive Bayes, Random Forest, KNN, AdaBoost, Multilayer Perceptron, Pipelining, and Ensemble Learning were some of the methods compared in Bagga et altechnique .'s proposed in 2020 [10] for the analysis of credit card fraud data. The methods and criteria used to identify fraud have an impact on how effectively it is done so.

Suman, Research Scholar, GJUS&T at Hisar HCE, introduced methods like supervised and unsupervised learning for credit card fraud detection in another study [12]. Even though some of these techniques and algorithms achieved unexpected success, they were unable to offer a reliable, long-lasting answer to fraud detection. Wen-Fang YU and Na Wang [11] described a related study area in which they employed outlier mining, outlier detection mining, and distance sum algorithms to accurately forecast fraudulent transaction in an emulation experiment using credit card transaction data set of one specific commercial bank. Data mining's field of outlier mining is primarily utilised in the financial and internet sectors. It focuses on identifying objects that are cut off from the primary system, such as transactions. They have measured the difference between the observed value of an attribute and its preset value by using attributes related to consumer behaviour and basing their calculations on those attributes' values. Unusual methods like hybrid data mining/complex network classification algorithms, which are based on network reconstruction algorithms and enable the creation of representations of the deviation of one instance from a reference group, have proven effective in the majority of medium-sized online transaction data sets.

III. METHODOLOGY

This model examines the dataset and identifies the fraudulent transactions using Python. In this study, a Kaggle dataset was evaluated. The dataset (creditcard.csv), which covers credit card transactions done by users across Europe in September 2013, is in CSV format and has a total of 284,807 transactions. Credit card transactions are split into two categories: fraudulent and non-fraudulent, depending on how the transaction behaves.

Only numerical input variables are used to present the Principal Component Analysis (PCA) transformation results. The primary components of PCA are Characteristics V1, V2,...V28; Time and Amount are the only features that remain unchanged. The seconds that elapsed between each transaction and the dataset's initial transaction are kept in the "Time" feature. The transaction Amount is represented by the feature "Amount," which can be utilised for example-dependent, cost-sensitive learning. When a fraudulent transaction is discovered, the 'Class' answer variable is evaluated to 1; otherwise, it is evaluated to 0.

The Local Outlier Factor (LOF) technique finds the anomalous data points by calculating the local deviation of a particular data point in relation to its neighbours.

To find outliers, this programme makes use of the local density. Locality is determined by the distance between the nearest neighbours, whereas density is determined by the distance between them. We can determine regions with a similar density and places with a denser area than their neighbours by comparing an object's local density to the local densities of its neighbours.

A data point is considered an outlier if its density is abnormally low when compared to its neighbours.
Outlier trends can be classified as either global or local. In contrast to a global outlier, which is an object that is distant from its k-th neighbour compared to the average, a local outlier is an object that is far from its neighbours' k-th nearest neighbours.

Data from a data frame can be used to make complex charts using ggplot2, a charting programme. It features a more programmable interface to specify which variables should be plotted, how they should be displayed, and other visual characteristics. As a result, if the underlying data changes or we want to switch from a bar plot to a scatterplot, we simply need to make a few minor tweaks. This makes it easier to generate charts that are suitable for publication with little modification and fine-tuning.

DESIGN
Feature creation
By translating the time to minutes or hours, the model's accuracy is raised.
Machine for gradient boosting (GBM): Confusion Matrix and several decision trees' predictions are merged to provide the final forecast.

IV. RESULTS AND DISCUSSION

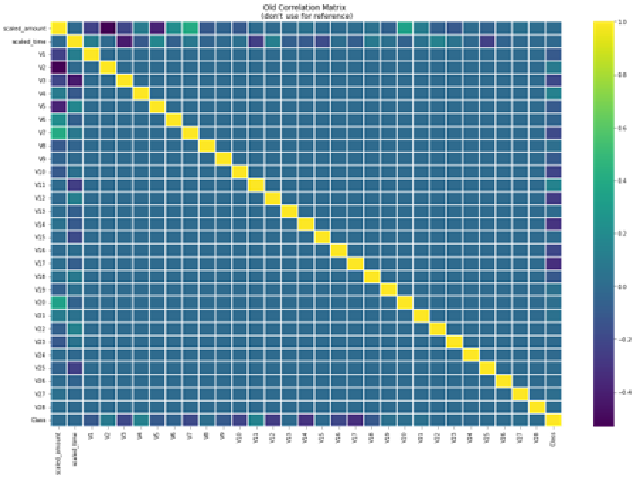

Figure 1. Heatmap of the original dataset

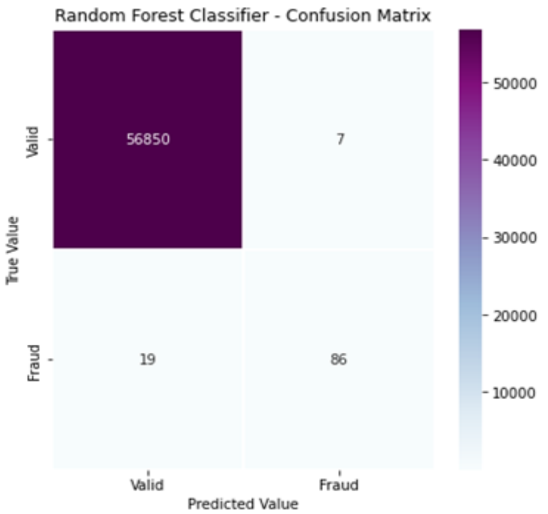

Figure 2.Confusion Matrix of the Random Forest Classifier with the accuracy of 0.96

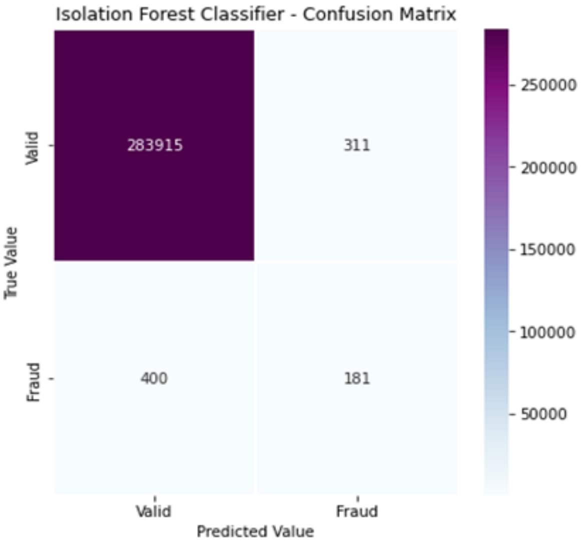

Figure 3. Confusion Matrix of the Isolation Forest Classifier with the accuracy of 0.68

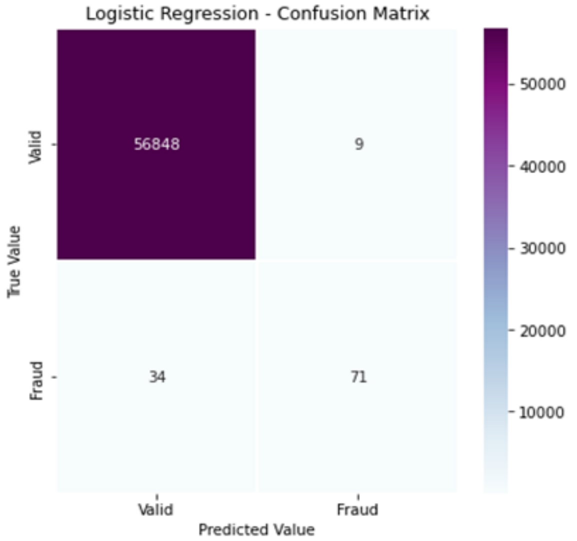

Figure 4. Confusion Matrix of the Logistic Regression with the accuracy of 0.89

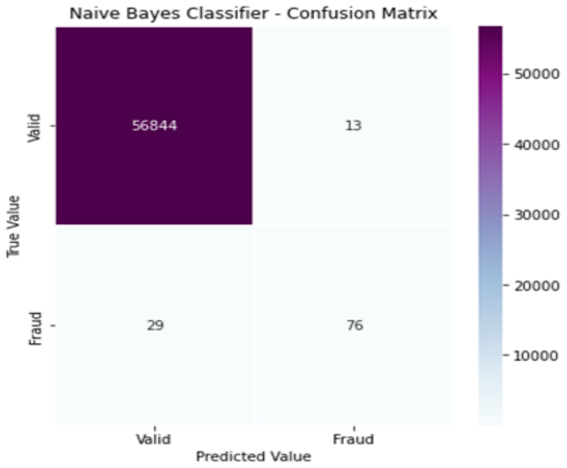

Figure 5.Confusion Matrix of the Naïve Bayes Classifier with the accuracy of 0.96

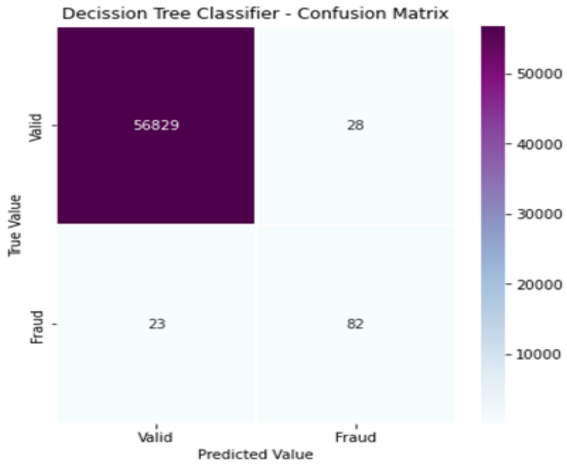

Figure 6.Confusion Matrix of the Decision Tree Classifier with the accuracy of 0.84

V. CONCLUSION AND FUTURE SCOPE

The possibility of credit card fraud is increasing as more people use credit cards for purchases. This study examines the detection of credit card fraud using machine learning methods like the decision tree and local outlier factor on a publicly accessible dataset. We also performed a normal distribution on the data, taking into account 492 fraud cases and 492 randomly created cases, in order to more effectively predict fraud transactions using boxplots. We got to the conclusion that the Random Forest Classifier offers the best accuracy of 0.96 after comparing the algorithms.

Despite not being able to reach our original goal of 100% accuracy in fraud detection, we were able to create a system that, given enough time and data, can come very close to it. As with any effort of this kind, there is room for improvement here. The project's structure makes it feasible to integrate numerous algorithms as modules and combine their outputs to increase the accuracy of the final result. This model can be improved even further by adding more algorithms. The output of these algorithms must, however, follow the same format as the others. Once that requirement is satisfied, as shown in the code, adding the modules is straightforward. This has a substantial amount of advantages. Due to this, the project is very flexible and adaptable. There are more growth potentials in the dataset. As was previously demonstrated, the algorithms' precision increases with dataset size. Thus, additional data will unquestionably increase the model's capacity to detect frauds and reduce the number of false positives. The banks themselves must, however, explicitly endorse this.

REFERENCES

[1]J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, “Credit card fraud detection using machine learning techniques: a comparative analysis,” in Proceedings of the 2017 International Conference on Computing Networking and Informatics (ICCNI), IEEE, Lagos, Nigeria, pp.**1–9, 2017.**

[2]A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi, and G. Bontempi, “Credit card fraud detection: realistic modeling and a novel learning strategy,” IEEE transactions on neural networks and learning systems, Vol.**29**, no.**8**, pp.**3784–3797, 2017.**

[3]S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang, and C. Jiang, “Random forest for credit card fraud detection,” in Proceedings of the 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), IEEE, Zhuhai, China, pp.**1–6**, **2018.**

[4]J. Jurgovsky, M. Granitzer, K. Ziegler et al., “Sequence classification for credit-card fraud detection,” Expert Systems with Applications, Vol.**100**, pp.**234–245, 2018.**

[5]D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, and A. Anderla, “Credit card fraud detection-machine learning methods,” in Proceeding of the 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH), IEEE, East Sarajevo, Bosnia and Herzegovina, March, pp.**1–5**, **2019,**

[6]F. Carcillo, Y.-A. Le Borgne, O. Caelen, Y. Kessaci, F. Oblé, and G. Bontempi, “Combining unsupervised and supervised learning in credit card fraud detection,” Information Sciences, Vol.**557**, pp.**317–331, 2021.**

[7]K. Randhawa, C. K. Loo, M. Seera, C. P. Lim, and A. K. Nandi, “Credit card fraud detection using AdaBoost and majority voting,” IEEE access, Vol.**6, 2018.**

[8]A. G. C. de Sá, A. C. M. Pereira, and G. L. Pappa, “A customized classification algorithm for credit card fraud detection,” Engineering Applications of Artificial Intelligence, Vol.**72**, pp.**21–29, 2018.**

[9]R. Sailusha, V. Gnaneswar, R. Ramesh, and G. R. Rao, “Credit card fraud detection using machine learning,” in Proceeding of the 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE, Madurai, India, pp.**1264–1270, 2020.**

[10]S. Bagga, A. Goyal, N. Gupta, and A. Goyal, “Credit card fraud detection using pipeling and ensemble learning,” Procedia Computer Science, Vol.**173**, pp.**104–112, 2020.**

[11]Wen-Fang Yu,Na Wang “Research on Credit Card Fraud Detection Model Based on Distance Sum” JCAI '09: Proceedings of the 2009 International Joint Conference on Artificial Intelligence April **2009.**

[12] Survey Paper on Credit Card Fraud Detection by Suman , Research Scholar, GJUS&T Hisar HCE, Sonepat published by International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Vol.**3** Issue.**3**, **2014.**

AUTHORS PROFILE

Sayee Chauhan is studying the bachelor's degree in Artificial Intelligence & Data Science from Vishwakarma Institute of Technology, Pune. Her research areas include machine learning, deep learning, machine learning, ensemble learning and operating systems.