International Journal of Innovative Research & Growth
A Peer Reviewed Journal
ISSN: 2455-1848
Journal homepage: http://www.ijirg.com/ijirg/
Volume: 10 Issue: 11, September 2021

Detection of credit card fraud transactions using machine learning algorithms techniques with data driven approaches: a comparative study

Shilpi Dandotia[1], Sandeep K. Tiwari[2*]
[1]Research Scholar, Vikrant Institute of Technology and Management, Gwalior, M. P., India
[2]Department of Computer Science and Engineering, Vikrant Institute of Technology and Management, Gwalior, M. P., India
E-mail: shilpidandotiaitm@gmail.com, sandeep72128@gmail.com
* Corresponding Author

Article Info
Received 15 July 2021
Received in revised form 10 August 2021
Accepted for publication 05 September 2021
DOI: 10.26671/IJIRG.2021.11.10.101
Citation:
Dandotia, S., Tiwari, S. K. (2021). Detection of credit card fraud transactions using machine learning algorithms techniques with data driven approaches: a comparative study. *Int J Innovat Res Growth*, 10, 56-61.

Abstract:
Finance fraud is a growing problem with far consequences in the financial industry and while many techniques have been discovered. It becomes challenging due to two major reasons–first, the profiles of normal and fraudulent behaviors change frequently and secondly due to reason that credit card fraud data sets are highly skewed. Credit card fraud events take place frequently and then result in huge financial losses [1]. The number of online transactions has grown in large quantities and online credit card transactions hold a huge share of these transactions. This paper focuses on four main fraud occasions in real-world transactions. Each fraud is addressed using a series of machine learning models and the best method is selected via an evaluation. This evaluation provides a comprehensive guide to selecting an optimal algorithm with respect to the type of the frauds and we illustrate the evaluation with an appropriate performance measure. Another major key area that we address in our project is real-time credit card fraud detection. For this, we take the use of predictive analytics done by the implemented machine learning models and an API module to decide if a particular transaction is genuine or fraudulent. This paper investigates and checks the performances of techniques are applied on the raw and preprocessed data. The performance of the techniques is evaluated based on accuracy, sensitivity, specificity and precision. In this paper, we will present a comparative study of some machine learning techniques, which gave the best results, according to our state of art [1] but applied to the same set of data. The objective of this study is to choose the best credit card fraud detection techniques to implement in our research work.
Keywords: - banking credit card fraud framework, machine learning techniques, artificial intelligence, etc.

1. Sources of Data

This review is based on the published academic articles as well as our RESEARCH REVIEW experience.

2. Introduction

Financial fraud is a growing concern with far reaching consequences in the government, corporate organizations, finance industry, In Today's world high dependency on internet technology has enjoyed increased credit card transactions but credit card fraud had also accelerated as online and offline transaction. As credit card transactions become a widespread mode of payment, focus has been given to recent computational methodologies to handle the credit card fraud problem. There are many fraud detection solutions and software which prevent frauds in businesses such as credit card, retail, e-commerce, insurance, and industries. Machine learning technique is one notable and popular methods used in solving credit fraud detection problem. It is impossible to be sheer certain about the true intention and rightfulness behind an application or transaction.

In reality, to seek out possible evidences of fraud from the available data using mathematical algorithms is the best effective option. Fraud detection in credit card is the truly the process of identifying those transactions that are fraudulent into two classes of legit class and fraud class transactions, several techniques are designed and implemented to solve to credit card fraud detection such as genetic algorithm, artificial neural network frequent item set mining, machine learning algorithms, migrating birds optimization algorithm, comparative analysis of logistic regression, SVM, decision tree and random forest is carried out Credit card fraud detection is a very popular but also a difficult problem to solve. Firstly, due to issue of having

only a limited amount of data, credit card makes it challenging to match a pattern for dataset. Secondly, there can be many entries in dataset with truncations of fraudsters which also will fit a pattern of legitimate behavior. Also the problem has many constraints. Firstly, data sets are not easily accessible for public and the results of researches are often hidden and censored, making the results inaccessible and due to this it is challenging to benchmarking for the models built. Datasets in previous researches with real data in the literature is nowhere mentioned. Secondly, the improvement of methods is more difficult by the fact that the security concern imposes a limitation to exchange of ideas and methods in fraud detection, and especially in credit card fraud detection. Lastly, the data sets are continuously evolving and changing making the profiles of normal and fraudulent behaviors always different that is the legit transaction in the past may be a fraud in present or vice versa. This paper evaluates four advanced data mining approaches, Decision tree, support vector machines, Logistic regression and random forests and then a collative comparison is made to evaluate that which model performed best.
Credit card transaction datasets are rarely available, highly imbalanced and skewed. Optimal feature (variables) selection for the models, suitable metric is most important part of data mining to evaluate performance of techniques on skewed credit card fraud data. A number of challenges are associated with credit card detection, namely fraudulent behavior profile is dynamic, that is fraudulent transactions tend to look like legitimate ones, Credit card fraud detection performance is greatly affected by type of sampling approach used, selection of variables and detection technique used.
In the end of this paper, conclusions about results of machine learning classifier evaluative testing are made and determined after getting review analysis.

3. Credit Card Fraud: Define

Credit card fraud is the most common and popular kind of identity theft and makes up 35.4% of all identity theft reports. Further, Identity theft occurs when someone uses information about you: name, birthday, social security number, bank statements, etc. to access your records or open new ones under your name. Credit and debit card fraud is a type of identity theft.
Here are some stats about identity theft:

- In 2018, Identity theft was the 3rd biggest cause of all categories of financial fraud in the USA, just below Imposter Scams and Debt and Loan Collection fraud.
- Identity theft makes up 14.8 percent of consumer fraud complaints with reports of 444,602 reported cases in 2018.

Credit card fraud was ranked #1 kind of Identity theft fraud - accounting for 35.4 percent of all identity theft fraud in 2018
Using identity information, creation of new accounts is up 24% from 2017 Takeover of existing accounts has decreased 6 percent from 2017.

4. Machine Learning Techniques: Fraud Detection Prudent Analysis

Machine learning and advancement of new technology Artificial intelligence is a powerful technology that has made its way into many fields like science, engineering, commerce and industry for dealing with massive datasets. It can be used to support a wide range of business intelligence applications such as customer profiling, target marketing, workflow management, store layout, and fraud detection. It allows users to analyses data from many different dimensions.
Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.
Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.
Currently, machine learning has been used in multiple fields and industries. For example, medical diagnosis, image processing, prediction, classification, regression etc.

5. Machine Learning: Prediction and Detection

Machine learning is the innovative inbuilt functionality based technology for detection of predictive tasks at any given problems. The training and test data set perform in the simulation of supervised algorithm for prediction of various banking fraud like credit card detection, cybercrime attacks detection etc. the supervised algorithm perform analysis on the basis of pre define data set in which perform prediction of any specific problems.

6. Classification

Classification is the method of finding a model or function that describes and differentiates data classes or concepts. The model is based on the analysis of a set of training data i.e., data objects for which the class labels are known. The model is used to predict the class label of objects for which the class label is unknown. Classification has number of applications, including fraud detection, target marketing, performance prediction and medical diagnosis.
Data classification is a two-step process, consisting of a learning step where a classification model is build and a classification step where the model is used to predict class labels for the given data.

- The Data Classification process consists of two steps: -
- Building the Classifier or Model.
- Using Classifier for Classification.

Machine Learning Is Sub-Categorized To Three Types

- Supervised Learning – Train and testing data set applied.
- Unsupervised Learning – clustering and association techniques applied.
- Reinforcement Learning – Hit & Trial.

7. Literature Review, SLR

The systematic literature review (SLR) on detection of credit card fraud transaction using machine learning techniques along with real time data driven approaches. In this paper we have mentioned various authors review on machine learning algorithm applied on detection of financial fraud detection with result and findings. The accuracy, efficiency of various classification and classifier algorithms also mentioned with perceptive values. The review findings also describing the data set which is real time driven or sample data set performed by various another researcher.

In [01] this paper author represents a research about a case study involving credit card fraud detection, where data normalization is applied before Cluster Analysis and with results obtained from the use of Cluster Analysis and Artificial Neural Networks on fraud detection has shown that by clustering attributes neuronal inputs can be minimized. And promising results can be obtained by using normalized data and data should be MLP trained. This research was based on unsupervised learning. Significance of this paper was to find new methods for fraud detection and to increase the accuracy of results.

In [02] In this paper author state that a new collative comparison measure that reasonably represents the gains and losses due to fraud detection is proposed. A cost sensitive method which is based on Bayes minimum risk is presented using the proposed cost measure. Improvements up to 23% are obtained when this method and other state of art algorithms are compared. The data set for this paper is based on real life transactional data by a large European company and personal details in data is kept confidential; accuracy of an algorithm is around 50%. Significance of this paper was to find an algorithm and to reduce the cost measure. The result obtained was by 23% and the algorithm they find was Bayes minimum risk.

In [03] this paper author explained that various modern techniques based on Sequence Alignment, Machine learning, Artificial Intelligence, Genetic Programming, Data mining etc. has been evolved and is still evolving to detect fraudulent transactions in credit card. A sound and clear understanding on all these approaches is needed that will certainly lead to an efficient credit card fraud detection system. Survey of various techniques used in credit card fraud detection mechanisms has been shown in this paper along with evaluation of each methodology based on certain design criteria. Analysis on Credit Card Fraud Detection Methods has been done. The survey in this paper was purely based to detect the efficiency and transparency of each method. Significance of this paper was conducting a survey to compare different credit card fraud detection algorithm to find the most suitable algorithm to solve the problem.

In [04] in this paper author execute A comparison has been made between models based on artificial intelligence along with general description of the developed fraud detection system are given in this paper such as the Naive Bayesian Classifier and the model based on Bayesian Networks, the clustering model. And in the end conclusions about results of models' evaluative testing are made. Number of legal truncations was determined greater or equal to 0.65 that is their accuracy was 65% using Bayesian Network. Significance of this paper is to compare between models based on artificial intelligence along with general description of the developed system and to state the accuracy of each model along with the recommendation to make the better model.

In [05] Nutan and Suman on review on credit card fraud detection they have supported the theory of what is credit card fraud, types of fraud like telecommunication, bankruptcy fraud etc. and how to detect it, in addition to it they have explained numerous algorithms and methods on how to detect fraud using Glass Algorithm, Bayesian, networks, Hidden Markova model, Decision Tree and 4 more. They have explained in detail about each algorithm and how this algorithm works along with mathematical explanation. Types of machine learning along with classifications have been studied. Pros and cons of each method is listed.

Many theories have been proposed to explain on how to detect credit card frauds using different methods, but Mohamad Zamini& Gholamali Montazer [06] had focused on deep auto encoder and k-means clustering as an unsupervised method to detect fraud by using and testing on 284807 transactions from European banks. Hypothesis for their research was imbalance data characteristics, time dependence between samples, concept drift and real-time detection. Further, as per the requirements of banks, they need to aim to catch maximum fraud transactions and lower the FNR simultaneously with reasonable timing. As evaluation measures, authors have used Tensor Flow in python and SKLearn library – an open-source. The accuracy of this method was 98.9%, as well as 81% TPR which outperforms in comparison with others. Although the literature presents a substantial increase in accuracy, this paper primarily focuses on the hybrid method of fraud detection.

Author Apapan Pumsirirat, Liu Yan [07] proposed a model of deep Auto-encoder and restricted Boltzmann machine (RBM) that can reproduce ordinary exchanges to discover oddities from typical examples. They have utilized deep learning dependent on auto-encoder (AE) is an unsupervised learning algorithm that applies back propagation by setting the sources of input equivalent to outputs. The RBM has two layers, the information layer (obvious) and a hidden layer. Likewise, they utilized the Tensor Flow library from Google to execute AE, RBM, and H2O by utilizing deep learning. Their outcomes show the mean squared error (MSE), root mean squared error (RMSE), and territory under the curve (AUC).

In this paper, Ishan Sohony, Rameshwar Pratap, Ullas Nambiar [08] proposed an ensemble method – based on a combination of random forest and feed-forward neural network. They saw that Random Forest is increasingly exact in distinguishing normal instances, and Neural Network is for identifying fraud instances. Along these lines, they considered a majority approach to only classify a transaction as fraud or normal if a majority of the classifiers classify it as fraud or

normal respectively. Consequently, we take the best of both universes and get a classifier whose precision and recall is optimal. Their experimental results point to this being a predominant method than other popular approaches.
Ogwueleka, Francisca Nonyelum [09] has highlighted the credit card fraud detection system that uses both conventional data mining and neural network approaches to deal with accomplish cooperative energy that better handles the Nigerian Credit Card fraud circumstance utilizing the four groups rather than two-phase model/cluster ordinarily utilized in the fraud detection system.). The author has revealed that receiver-operating curve (ROC) for credit card fraud (CCF) detection watch distinguished over 95% of fraud cases without causing bogus alerts, not at all like other factual models and the two-phase clusters.
In research introduced by Maes, S., Tuyls, K., Vanschoenwinkel, B., & Manderick, B. [10] to identify Credit Card Fraud issue, they tried utilizing two AI methods individually Artificial Neural Network (ANN) and Bayesian Belief Network (BBN). Author have detailed outcome with ANN and BNN, for a False Positive Rate of separately 10% and 15%. What's more, they reasoned that BNN gives better outcomes concerning fraud detection and their training period is shorter yet fraud detection process significantly quicker with ANN.
Authors [11] presented Real Time Data-Driven Approaches for Credit Card Fraud Detection. They only focused on the one-class classification methods for anomaly detection. Anomaly detection is a method to identify whether or not a metric is behaving differently than it has in the past, taking into account trends. This paper fundamentally centered around one class support vector machine (OCSVM) with the optimal kernel parameter selection and T2 control graph. As execution measurements, creators have utilized DR, FPR, precision, and F-score. What's more, creators have inferred that the proposed approach accomplished a significant level of discovery precision and a low false alarm rate. Their methodologies will bring numerous advantages for the associations and for singular clients regarding cost and time proficiency.

8. Suggestion and Findings

This research is to detect the credit card fraud in the dataset obtained from real time data driven by applying Logistic regression, Decision tree, SVM, Random Forest, Data driven approaches, tensor flow method, deep learning, CNN convolutional neural network and to evaluate their Accuracy, sensitivity, specificity, precision using different models and compare and collate them to state the best possible model to solve the credit card fraud detection problem.
Ability of system to automatically learn and improve from experience without being explicitly programmed is called machine learning and it focuses on the development of computer programs that can access data and use it learn for themselves. And classifier can be stated as an algorithm that is used to implement classification especially in concrete implementation, it also refers to a mathematical function implemented by algorithm that will map input data into category. It is an instance of supervised learning i.e. where training set of correctly identified observations is available.
First the credit card dataset is taken from the source and cleaning and validation is performed on the dataset which includes removal of redundancy, filling empty spaces in columns, converting necessary variable into factors or classes then data is divided into 2 part, one is training dataset and another one is test data set. Now K fold cross validation is done that is the original sample is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k −1 subsamples are used as training data, Models are created for machine learning techniques and artificial intelligence and then accuracy, sensitivity, specificity, precision are calculated and a comparison is made.

9. Comparison Analysis

01. Logistic Regression is a supervised classification method that returns the probability of binary dependent variable that is predicted from the independent variable of dataset that is logistic regression predict the probability of an outcome which has two values either zero or one, yes or no and false or true.
02. SVM is a one of the popular machine learning algorithm for regression, classification. It is a supervised learning algorithm that analyses data used for classification and regression. SVM modeling involves two steps, firstly to train a data set and to obtain a model & then, to use this model to predict information of a testing data set.
03. Decision tree is an algorithm that uses a tree like graph or model of decisions and their possible outcomes to predict the final decision, this algorithm uses conditional control statement. A Decision tree is an algorithm for approaching discrete-valued target functions, in which decision tree is denoted by a learned function. For inductive learning these types of algorithms are very famous and have been successfully applied to abroad range of tasks.
04. Random Forest is an algorithm for classification and regression. Summarily, it is a collection of decision tree classifiers. Random forest has advantage over decision tree as it corrects the habit of over fitting to their training set. A subset of the training set is sampled randomly so that to train each individual tree and then a decision tree is built; each node then splits on a feature selected from a random subset of the full feature set.

METRICES	LOGISITC REGRESION	SVM	DECISION TREE	RANDOM FOREST
ACCURACY	0.97	0.975	0.955	0.986
SENSITIVITY	0.975	0.973	0.955	0.984
SPECIFICITY	0.92	0.912	0.877	0.904
PRECISION	0.96	0.996	0.995	0.997

Table: 1 Comparative result analysis by performance matrices+

Table 1.2 Confusion matrix format: define

S.NO	ACTUAL /PREDICTED	NOT FRAUD DETECTION	FRAUD DETECTION
01	FRAUD DETECT	TRUE POSITIVE	FALSE
02	FRAUD NOT DETECT	FALSE NEGATIVE	TRUE NEGATIVE

10. Conclusion

From the experiments the result that has been concluded is that Logistic regression has a accuracy of 97.7% while SVM shows accuracy of 97.5% and Decision tree shows accuracy of 95.5% but the best results are obtained by Random forest with a precise accuracy of 98.6%. The results obtained thus conclude that Random forest shows the most precise and high accuracy of 98.6% in problem of credit card fraud detection with dataset provided by machine learning. This paper conclude various comparative analysis result determined using authors review conclusion and findings. The various algorithms of machine learning worked on detection of credit card fraud with real time data driven. To overcome this problem we will use in our research work the tensor flow method and new machine learning approaches with real data driven of credit card transaction implement.

In our research we will perform the detection of credit card fraud using python coding along with Jupiter navigator simulation tool. The result will show the new approaches for real time data driven data set and determine performance matrices will greater than other machine learning previous algorithms.

References

i. Raj, S. B. E. and Portia, A. A. (2011). Analysis on credit card fraud detection methods, Computer, Communication and Electrical Technology. International Conference on (ICCCET), 152-156.

ii. Jain, R., Gour, B. and Dubey, S. (2016). A hybrid approach for credit card fraud detection using rough set and decision tree technique. International Journal of Computer Applications, 139, 1-6.

iii. Dermala, N. and Agrawal, A. (2016). Credit card fraud detection using SVM and Reduction of false alarms. International Journal of Innovations in Engineering and Technology, 7, 176-182.

iv. Phua, C., Lee, V., Smith, K. and Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. arXiv preprint arXiv:1009.6119.

v. Zamini, M. and Montazer, G. (2018). Credit card fraud detection using auto encoder based clustering. 9th International Symposium on Telecommunications (IST). IEEE.

vi. Pumsirirat, Apapan, and Yan, L. (2018). Credit card fraud detection using deep learning based on auto-encoder and restricted boltzmann machine. International Journal of advanced computer science and applications, 9, 18-25.

vii. Song C., Liu F., Huang Y., Wang L. and Tan T. (2013). Auto-encoder Based Data Clustering. In: Ruiz-Shulcloper J., Sanniti di Baja G. (eds) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2013. Lecture Notes in Computer Science, 8258. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-41822-8_15

viii. Yuan, S., Wu, X., Li, J. and Lu, A. (2017). Spectrum-based deep neural networks for fraud detection. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management.

ix. Sohony, I., Pratap, R. and Nambiar, U. (2018). Ensemble learning for credit card fraud detection. Proceedings of the ACM India Joint International Conference on Data Science and Management of Data, 289-294. https://doi.org/10.1145/3152494.3156815

x. Zou, J., Zhang, J. and Jiang, P. (2019). Credit Card Fraud Detection Using Autoencoder Neural Network. arXiv preprint arXiv:1908.11553.

xi. Ogwueleka, F. N. (2011). Data mining application in credit card fraud detection system. Journal of Engineering Science and Technology, 6, 311-322.

xii. Maes, S., Tuyls, K., Vanschoenwinkel, B. and Manderick, B. (2002). Credit Card Fraud Detection Using Bayesian and Neural Networks. In Proceedings of the First International NAISO Congress on NEURO FUZZY THECHNOLOGIES

January 16 - 19, 2002 (Havana, Cuba), Proceedings of the First International NAISO Congress on NEURO FUZZY THECHNOLOGIES January 16 - 19, 2002 (Havana, Cuba)
xiii. Koziarski, M. and Woźniak, M. (2017). CCR: A combined cleaning and resampling algorithm for imbalanced data classification. International Journal of Applied Mathematics and Computer Science, 27, 727-736. https://doi.org/10.1515/amcs-2017-0050
xiv. H. Najadat, Altiti, O.A., Aqouleh, A. A. and Younes, M. (2020). Credit Card Fraud Detection Based on Machine and Deep Learning. 2020 11th International Conference on Information and Communication Systems (ICICS), 2020, 204-208, doi: 10.1109/ICICS49469.2020.239524.
xv. Zareapoor, M. and Shamsolmoali, P. (2015). Application of credit card fraud detection: Based on bagging ensemble classifier. International Conference on Intelligent Computing, Communication & Convergence (ICCC-2014), Procedia computer science 48.2015 (2015): 679-686.
xvi. Tran, P. H., Tran, K. P., Huong, T. T., Heuchenne, C. et al. (2018). Real time data-driven approaches for credit card fraud detection. Proceedings of the 2018 International Conference on E-Business and Applications, 6-9. https://doi.org/10.1145/3194188.3194196
xvii. Seeja, K. R., and Zareapoor, M. (2014). Fraud Miner: A novel credit card fraud detection model based on frequent itemset mining. The Scientific World Journal. 2014. https://doi.org/10.1155/2014/252797
xviii. Roy, A., Sun, J., Mahoney, R., Alonzi, L., Adams, S. and Beling, P.A. (2018). Deep learning detecting fraud in credit card transactions. 2018 Systems and Information Engineering Design Symposium (SIEDS), 129-134, doi: 10.1109/SIEDS.2018.8374722.
xix. Dighe, D., Patil, S. and Kokate, S. (2018). Detection of Credit Card Fraud Transactions Using Machine Learning Algorithms and Neural Networks: A Comparative Study. 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 1-6.
xx. Puh, M. and Brkić, L. (2019). Detecting Credit Card Fraud Using Selected Machine Learning Algorithms. 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 1250-1255.
xxi. Awoyemi, J.O., Adetunmbi, A.O. and Oluwadare, S.A. (2017). Credit card fraud detection using machine learning techniques: A comparative analysis. 2017 International Conference on Computing Networking and Informatics (ICCNI), 1-9.
xxii. Pillai, T. R., Hashem, I. A. T., Brohi, S. N., Kaur, S. and Marjani, M. (2018). Credit Card Fraud Detection Using Deep Learning Technique. 2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA), 2018, 1-6, doi: 10.1109/ICACCAF.2018.8776797.
xxiii. Kazemi, Z. and Zarrabi, H. (2017). Using deep networks for fraud detection in the credit card transactions. 2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI), 0630-0633.
xxiv. de Sá, A. G. C., Pereira, A. C. M. and Pappa, G. L. (2018). A customized classification algorithm for credit card fraud detection. Engineering Applications of Artificial Intelligence, 72, 21-29. https://doi.org/10.1016/j.engappai.2018.03.011