IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 1, January 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.512

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 1, January 2021 ||

DOI:10.15680/IJIRSET.2021.1001017

Card Fraud Detection Using Machine Learning

Saurabh Kumar Lall[1], Pranav Pranjal[2], Sarabjeet Kumar[3], Mrs. Mayuri KP[4]

Department of Computer Science and Engineering, Sir M.Visvesvaraya Institute of Technology, Bangalore, India[1,2,3,4]

ABSTRACT: The detection of frauds in credit card transactions is a major topic in financial research, of profound economic implications. While this has hitherto been tackled through data analysis techniques, the resemblances between this and other problems, like the design of recommendation systems and of diagnostic/prognostic medical tools, suggest that a complex network approach may yield important benefits. In this paper we present a first hybrid data mining/complex network classification algorithm, able to detect illegal instances in a real card transaction data set. It is based on a recently proposed network reconstruction algorithm that allows creating representations of the deviation of one instance from a reference group. We show how the inclusion of features extracted from the network data representation improves the score obtained by a standard, neural network-based classification algorithm and additionally how this combined approach can outperform a commercial fraud detection system in specific operation niches. Beyond these specific results, this contribution represents a new example on how complex networks and data mining can be integrated as complementary tools, with the former providing a view to data beyond the capabilities of the latter.

KEYWORDS: Credit card fraud, applications of machine learning, isolation forest algorithm, local outlier factor, CNN Model, Heat map.

I.INTRODUCTION

In daily routine we use credit cards to buy goods and services using online transaction or physical card for offline transaction In credit card based purchase, the card holder issues his card to merchant to do payment .the person has to steal the card to make the transaction fraudulent . If the user is not aware of loss of card it leads to financial loss to the user as well as credit card company. When the payment mode is online, attackers require only little information for doing false transaction. Example card number. The only way to detect these kind of fraud is to analyse the spending patterns on every card and irregularities are figured with respect to normal pattern. Fraud which is detected using existing purchase data of card holder is way to reduce the rate of frauds. Every card holder is characterised by patterns containing information about distinctive purchase category the time since the last buying, money spent and other things.
'Fraud' in credit card transactions is unauthorized and unwanted usage of an account by someone other than the owner of that account. Necessary prevention measures can be taken to stop this abuse and the behavior of such fraudulent practices can be studied to minimize it and protect against similar occurrences in the future. In other words, Credit Card Fraud can be defined as a case where a person uses someone else's credit card for personal reasons while the owner and the card issuing authorities are unaware of the fact that the card is being used.

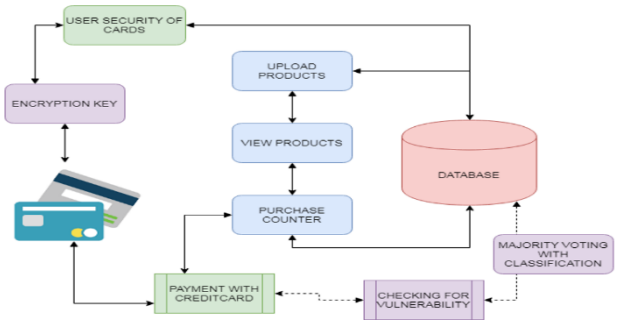
Also, the transaction patterns often change their statistical properties over the course of time. These are not the only challenges in the implementation of a real-world fraud detection system, however. In real world examples, the massive stream of payment requests is quickly scanned by automatic tools that determine which transactions to authorize.
Machine learning algorithms are employed to analyse all the authorized transactions and report the suspicious ones. These reports are investigated by professionals who contact the cardholders to confirm if the transaction was genuine or fraudulent.

The investigators provide a feedback to the automated system which is used to train and update the algorithm to eventually improve the fraud-detection performance over time.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 1, January 2021 ||

DOI:10.15680/IJIRSET.2021.1001017



Some of the currently used approaches to detection of such fraud are:

- Artificial Neural Network
- Fuzzy Logic
- Genetic Algorithm
- Logistic Regression
- Decision tree
- Bayesian Networks
- Hidden Markov Model
- K-Nearest Neighbour

II. LITERATURE REVIEW

Credit card fraud detection has received significant consideration from researchers in the world. Several techniques have been developed to detect fraud using credit card which are based on Bayesian networks, neural network, data mining, clustering techniques, genetic algorithms, decision tree etc. The idea of similarity tree, a variety of Decision Trees logic was proposed by Kokkinaki in 1997. A decision tree is defined recursively; it contains nodes and edges that are labeled with attribute names and with values of attributes, respectively. All of these satisfy some condition and get an intensity factor which is defined as the ratio of the number of transactions that satisfy applied conditions over the total number of legitimate transaction.

The advantages of the method are the ease to understand and implement. While its disadvantages include a need for continual updates when user habits and fraud patterns change, as the user profiles are not dynamically adaptive. Ghosh and Reilly proposed a neural network method to detect credit card fraud transactions. They built a detection system, which uses a three-layered feed-forward network with only two training passes and is trained on a large sample of labeled credit card account transactions. These transactions contains sample fraud cases due to stolen cards, lost cards, application fraud, stolen card details, counterfeit fraud etc. They tested on a data set of all transactions of credit card account over a subsequent period of time. The system significantly reduced the investigation workload of fraud analysts.

III. PROBLEM STATEMENT

Our goal is to implement 3 different machine learning models in order to classify, to the highest possible degree of accuracy, credit card fraud from a dataset gathered in Europe in 2 days in September 2013. After initial data exploration, we knew we would implement a logistic regression model, a k-means clustering model, and a neural network. Enormous Data is processed every day and the model build must be fast enough to respond to the scam in time.Contrary to popular belief, merchants are far more at risk from credit cardfraud than the cardholders. While consumers may face trouble trying to get afraudulent charge reversed, merchants lose the cost of the product sold, paychargeback fees, and fear from the risk of having their merchant account closed.

Some challenges we observed from the start were thehuge imbalance in the dataset: frauds only account for 0.172% of fraud transactions. In this case, it is much worse to have false negatives than false positives in our predictions because false negatives mean that someone gets away with credit card fraud. False positives, on the other hand, merely cause a

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 1, January 2021 ||

DOI:10.15680/IJIRSET.2021.1001017

complication and possible hassle when a cardholder must verify that they did, in fact, complete said transaction (and not a thief).

IV.PROPOSED SYSTEM

In this paper, we have used the random forest and isolation forest algorithms are used for detecting credit card fraud. The algorithms range from standard neural networks to deep learning models. They are evaluated using both benchmark and real-world credit card data sets. To further evaluate the robustness and reliability of the models, noise is added to the real-world data set. The key contribution of this paper is the evaluation of a variety of machine learning models with a real-world credit card data set for fraud detection. While other researchers have used various methods on publicly available data sets, the data sets used in this paper are extracted from actual credit card transaction information over three months.

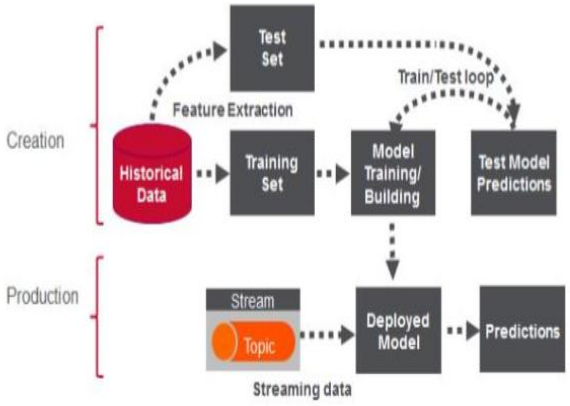


Figure. Project Flow diagram

A class's prior may be calculated by assuming equiprobable classes (i.e., priors = 1 / (number of classes)), or by calculating an estimate for the class probability from the training set (i.e., (prior for a given class) = (number of samples in the class) / (total number of samples)). To estimate the parameters for a feature's distribution, one
must assume a distribution or generate nonparametric models for the features from the training set. [8] The assumptions on distributions of features are called the event model of the Naive Bayes classifier. For discrete features like the ones encountered in document classification (include spam filtering), multinomial and Bernoulli distributions are popular.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 1, January 2021 ||

DOI:10.15680/IJIRSET.2021.1001017

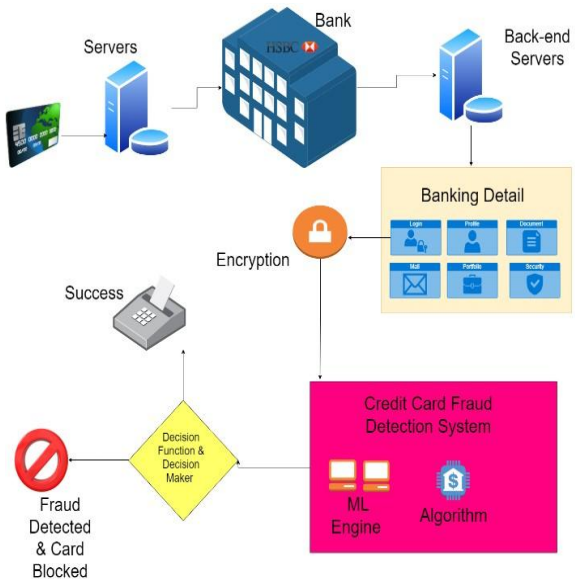
These assumptions lead to two distinct models, which are often confused. It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

V. METHODOLOGY

The approach that this paper proposes, uses the latest machine learning algorithms to detect anomalous activities, called outliers.

The full architecture diagram can be represented as follows:



First of all, we obtained our dataset let's separate the Fraudulent cases from the authentic ones and compare their occurrences in the dataset as follows:

After this analysis, we plot a heatmap to get a coloured representation of the data and to study the correlation between out predicting variables and the class variable.

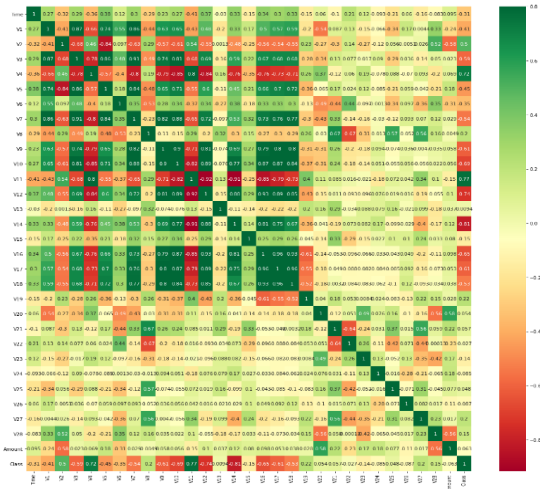
International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

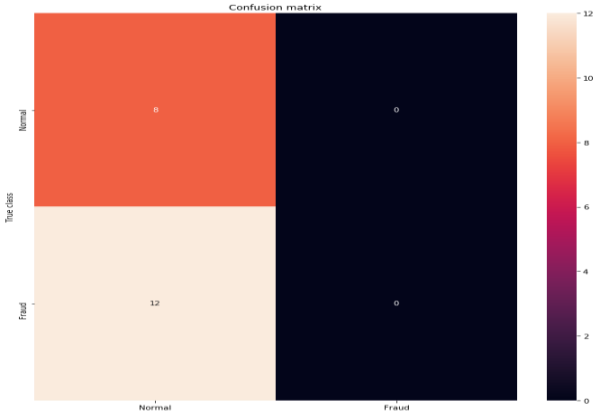
|| Volume 10, Issue 1, January 2021 ||

DOI:10.15680/IJIRSET.2021.1001017

This heatmap is shown below:



The confusion matrix between the normal and fraud transaction shown below:



The dataset is now formatted and processed. The time and amount column are standardized and the Class column is removed to ensure fairness of evaluation. The data is processed by a set of algorithms from modules. The following module diagram explains how these algorithms work together: This data is fit into a model and the Random forest classifier are applied on it.

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

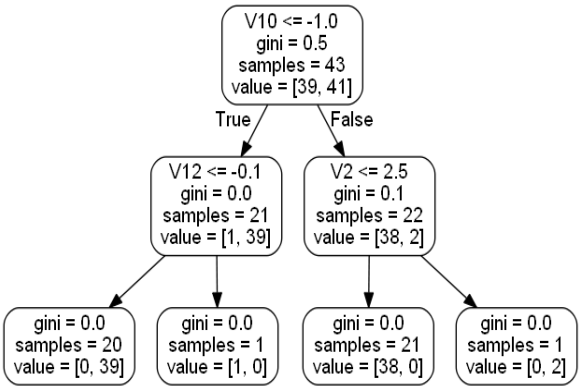
|| Volume 10, Issue 1, January 2021 ||

DOI:10.15680/IJIRSET.2021.1001017

The pseudocode for this algorithm is written as:

```
#visualizing the random tree
feature_list = list(X.columns)
#Import tools needed for visualization
from IPython.display import Image
from sklearn.tree import export_graphviz
import pydot
#pulling out one tree from the forest
tree = rfc.estimators_[5]
export_graphviz(tree, out_file = 'tree.dot', fe
ature_names = feature_list, rounded = True, pre
cision = 1)
# Use dot file to create a graph
(graph, ) = pydot.graph_from_dot_file('tree.do
t')
# Write graph to a png file
display(Image(graph.create_png()))
```

Plotting the results of visualizing the random tree:



VI. RESULT ANALYSIS

The algorithm used here is Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

```
# Building the Random Forest Classifier (RANDOM FOREST)
from sklearn.ensemble import RandomForestClassifier
# random forest model creation
rfc = RandomForestClassifier()
rfc.fit(X_train,Y_train)
# predictions
y_pred = rfc.predict(X_test)
```

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 1, January 2021 ||

DOI:10.15680/IJIRSET.2021.1001017

Random decision forests correct for decision trees' habit of overfitting to their training set which gives the following results:

```
The model used is Random Forest classifier
The accuracy is 0.95
The precision is 1.0
The recall is 0.9166666666666666
The F1-Score is 0.9565217391304348
The Matthews correlation coefficient is0.90267093384844
```

VII. CONCLUSION

Credit card fraud is without a doubt an act of criminal dishonesty. This article has listed out the most common methods of fraud along with their detection methods and reviewed recent findings in this field. This paper has also explained in detail, how machine learning can be applied to get better results in fraud detection along with thealgorithm, pseudocode, explanation its implementation and experimentation results.

While the algorithm does reach over 99.6% accuracy, its precision remains only at 28% when a tenth of the data set is taken into consideration. However, when the entire dataset is fed into the algorithm, the precision rises to 33%. This high percentage of accuracy is to be expected due to the huge imbalance between the number of valid and number of genuine transactions.

Since the entire dataset consists of only two days' transaction records, its only a fraction of data that can be made available if this project were to be used on a commercial scale. Being based on machine learning algorithms, the program will only increase its efficiency over time as more data is put into it.

VIII. FUTURE ENHANCEMENTS

While we couldn't reach out goal of 100% accuracy in fraud detection, we did end up creating a system that can, with enough time and data, get very close to that goal. As with any such project, there is some room for improvement here. The very nature of this project allows for multiple algorithms to be integrated together as modules and their results can be combined to increase the accuracy of the final result. This model can further be improved with the addition of more algorithms into it. However, the output of these algorithms needs to be in the same format as the others. Once that condition is satisfied, the modules are easy to add as done in the code. This provides a great degree of modularity and versatility to the project.

More room for improvement can be found in the dataset. As demonstrated before, the precision of the algorithms increases when the size of dataset is increased. Hence, more data will surely make the model more accurate in detecting frauds and reduce the number of false positives. However, this requires official support from the banks themselves.

Further enhancement can be done by making this system secure with the useof certificates for both merchant and customer and as technology changes new checks

can be added to understand the pattern of fraudulent transactions and to alert therespective card holders and bankers when fraud activity is identified.

REFERENCES

[1] A. C. Bahnsen, D. Aouada, A. Stojanovic, and B. Ottersten, ―Detecting credit card fraud using periodic features,‖ in Proc. 14th Int. Conf. Mach. Learn. Appl., Dec. 2015, pp. 208–213

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[3] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[4] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] I.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271350.

[7] G. Eason, B. Noble, and I.N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.512

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details