ISSN (Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 4, April 2019

Credit Card Fraud Detection Using Random Forest Technique

B. Mohankumar[1], Dr. K. Karuppasamy[2]

Assistant Professor, Department of Information Technology, Sri Ramakrishna Engineering College, Coimbatore, Tamil Nadu, India[1]

Professor & Head, Department of Computer Science and Engineering, RVS College of Engineering and Technology, Coimbatore, Tamil Nadu[2]

ABSTRACT*:* In our project, mainly focussed on credit card fraud detection for in real world. Collect the credit card datasets for trained dataset. Then will provide the user credit card queries for testing data set. After classification process of random forest algorithm using to the already analyzing data set and user provide current dataset. Optimizing the accuracy of the result data. Then will apply the processing of some of the attributes provided can find affected fraud detection in viewing the graphical model visualization. The performance of the techniques is evaluated based on accuracy, sensitivity, and specificity, precision. The results indicate about the optimal accuracy for Random Forest is 98.6% respectively

KEYWORDS: Credit card, Fraud Detection, Accuracy, Random Forest, Dataset.

I. INTRODUCTION

Financial fraud is increases in modern communication world within seconds. They stolen billion of dollars.tis way company and financial institutes are losses there profit, mainly all the bank transactions are now converted in online. In online we have username and password .so they provide credit card for purchasing and transaction purpose. Credit card is more relevant for day by day for a person. One person behaviour we understand followed by credit card usage. By detecting this fraud cases different software are developed but by chance it can not existing much more years. So we are going for next stage for detecting fraudulent cases in credit card by machine learning approach. Machine learning approach is based on algorithm performance, so here we use much accurate algorithm Random forest .this is the best algorithm for classification. These analysis has taken by choose different attributes of credit card. Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on models and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model of sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task Machine learning algorithms are used in the applications of email filtering, detection of network intruders, and computer vision, where it is infeasible to develop an algorithm of specific instructions for performing the task. Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a field of study within machine learning, and focuses on exploratory data analysis through unsupervised learning In its application across business problems, machine learning is also referred to as predictive analytics.

ISSN (Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 4, April 2019

II. DESCRIPTION OF MODULES

PRE PROCESSING:-

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Pre processing is a technique that is used to convert the raw data into a clean data set. In python, scikit-learn library has a pre-built functionality under sklearn, pre processing. There are many more options for pre-processing which we'll explore.

FEATURE EXTRACTION:-

In machine learning, pattern recognition and in image processing, feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations. Feature extraction is a dimensionality reduction process, where an initial set of raw variables is reduced to more manageable groups (features) for processing, while still accurately and completely describing the original data set. Text Analysis is a major application field for machine learning algorithms. However the raw data, a sequence of symbols cannot be fed directly to the algorithms themselves as most of them expect numerical feature vectors with a fixed size rather than the raw text documents with variable length. The main idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of many variables correlated with each other, either heavily or lightly, while retaining the variation present in the dataset, up to the maximum extent. The same is done by transforming the variables to a new set of variables, which are known as the principal components (or simply, the PCs) and are orthogonal, ordered such that the retention of variation present in the original variables decreases as we move down in the order. So, in this way, the 1st principal component retains maximum variation that was present in the original components. The principal components are the eigenvectors of a covariance matrix, and hence they are orthogonal.

Importantly, the dataset on which PCA technique is to be used must be scaled. The results are also sensitive to the relative scaling. As a layman, it is a method of summarizing data. Imagine some wine bottles on a dining table. Each wine is described by its attributes like colour, strength, age, etc. But redundancy will arise because many of them will measure related properties. So what PCA will do in this case is summarize each wine in the stock with less characteristics. Intuitively, Principal Component Analysis can supply the user with a lower-dimensional picture, a projection or "shadow" of this object when viewed from its most informative viewpoint. Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and computer science. Data analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, while being used in different business, science, and social science domains. In today's business, data analysis is playing a role in making decisions more scientific and helping the business achieve effective operation.

CLASSIFICATION:-

In machine learning and statistics, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. it is a technique where we categorize data into a given number of classes. Classification model: A classification model tries to draw some conclusion from the input values given for training. It will predict the class labels/categories for the new data. For classification we use machine learning algorithm .as well as prediction.

ISSN (Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 4, April 2019

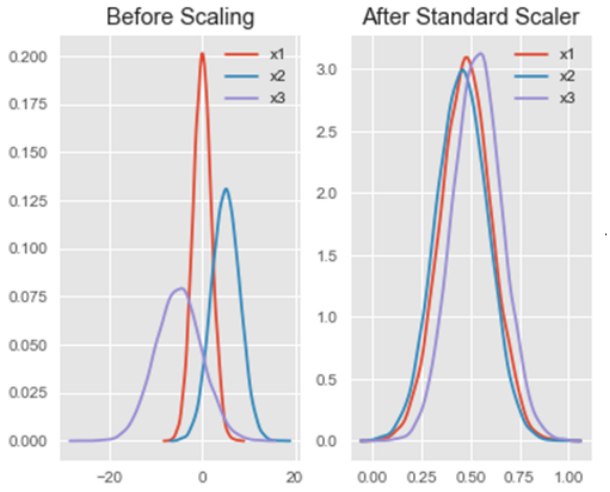
PREDICTION:-

Machine learning is a way of identifying patterns in data and using them to automatically make predictions or decisions. The two main methods of machine learning you will focus on are regression and classification. Here we predict fraudulent or non -fraudulent by algorithm performance.

III. BLOCK DIAGRAM

In this block diagram given dataset is splited in to tested and trained data. Training set is the one on which we train and fit our model basically to fit the parameters but test data is used only to assess performance of model. Training data is used to fit your model. Test data is used to validate your model. If you see a big performance difference, the model is not validated. This trained data is under go pre-processing .its nothing but alignment of data..In dataset contains random oriented data so we have to remove all the noisy data like not a number formate, spaces columns etc. Different pre processing technique are in machine learning, standard scalar is used here. The StandardScaler assumes your data is normally distributed within each feature and will scale them such that the distribution is now centred around 0, with a standard deviation of 1.The mean and standard deviation are calculated for the feature and then the feature is scaled based on:

xi–mean(x)/std dev(x).



In feature extraction we choose smote algorithm. SMOTE stands for Synthetic Minority Oversampling Technique. This is a statistical technique for increasing the number of cases in your dataset in a balanced way. The module works by generating new instances from existing minority cases that you supply as input. This implementation of SMOTE does **not** change the number of majority cases. After completion of pre- processing and feature extraction we going for algorithm performance. A classifier of random forest I, which is a simple implement of decision tree, is called a random tree . The training set of each tree is a collection of samples selected randomly from the standard training set with replacement. At each internal node, it randomly selects a subset of attributes and computes the centers of different classes of the data in current node. The centers of class 0 and 1 are denoted as Non –fraud and fraud, respectively. In Random forest the number of cases in the training set is N, sample N cases at random - but with replacement, from the original data. This sample will be the training set for growing the tree.If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.Each tree is grown to the largest extent possible.

ISSN (Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 4, April 2019

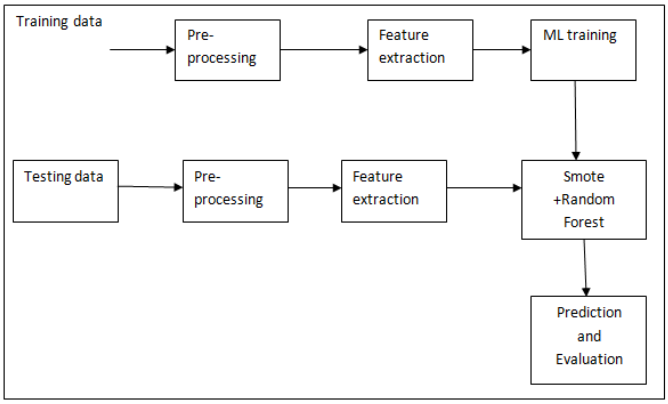


Fig. 1. Block diagram

IV. SOFTWARE DESCRIPTION

PYTHON:-

Python is an open source programming language. Python was made to be easy-to-read and powerful. A Dutch programmer named Guido van Rossum made Python in 1991. He named it after the television show Monty Python's Flying Circus. Many Python examples and tutorials include jokes from the show. Python is an interpreted language. Interpreted languages do not need to be compiled to run. A program called an interpreter runs Python code on almost any kind of computer. This means that a programmer can change the code and quickly see the results. This also means Python is slower than a compiled language like C, because it is not running machine code directly. Python is a good programming language for beginners. It is a high-level language, which means a programmer can focus on what to do instead of how to do it. Writing programs in Python takes less time than in some other languages. Python has a very easy-to-read syntax. Some of Python's syntax comes from C, because that is the language that Python was written in. But Python uses whitespace to delimit code: spaces or tabs are used to organize code into groups. This is different from C. In C, there is a semicolon at the end of each line and curly braces ({}) are used to group code. Using whitespace to delimit code makes Python a very easy-to-read language. Some things that Python is often used for are:

- Web development
- Game programming
- Desktop GUIs
- Scientific programming

LIBRARY MODULES:-

1. NUMPY:-

NumPy is a Python package. It stands for 'Numerical Python'. It is a library consisting of multidimensional array objects and a collection of routines for processing of array. **Numeric**, the ancestor of NumPy, was developed by Jim Hugunin. Another package Numarray was also developed, having some additional functionalities. In 2005, Travis Oliphant created NumPy package by incorporating the features of Numarray into Numeric package. There are many contributors to this open source project.

Using NumPy, a developer can perform the following operations:

- Mathematical and logical operations on arrays.
- Fourier transforms and routines for shape manipulation.
- Operations related to linear algebra. NumPy has in-built functions for linear algebra and random number generation.

ISSN (Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 4, April 2019

2. PANDAS:-

Pandas is an open-source, BSD-licensed Python library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc. In this tutorial, we will learn the various features of Python Pandas and how to use them in practice.

Pandas deals with the following three data structures :

- Series
- DataFrame
- Panel

3.MATPLOITLIB:-

It is a collection of command style functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc. In matplotlib.pyplot various states are preserved across function calls, so that it keeps track of things like the current figure and plotting area, and the plotting functions are directed to the current axes (please note that "axes" here and in most places in the documentation refers to the *axes* part of a figure and not the strict mathematical term for more than one axis).

4.SKLEARN:-

Scikit-learn is a machine learning library for Python. It features several regression, classification and clustering algorithms including SVMs, gradient boosting, k-means, random forests and DBSCAN. It is designed to work with Python Numpy and Scipy.The scikit-learn project kicked off as a Google Summer of Code (also known as GSoC) project by David Cournapeau as scikits.learn. It gets its name from “Scikit”, a separate third-party extension to

V. EXPERIMENTAL RESULTS

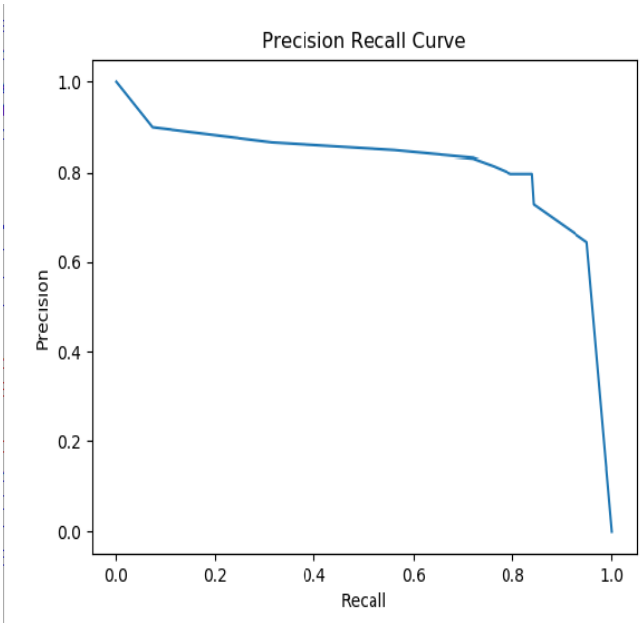


Fig. 2. Precision Recall Curve

ISSN (Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 4, April 2019

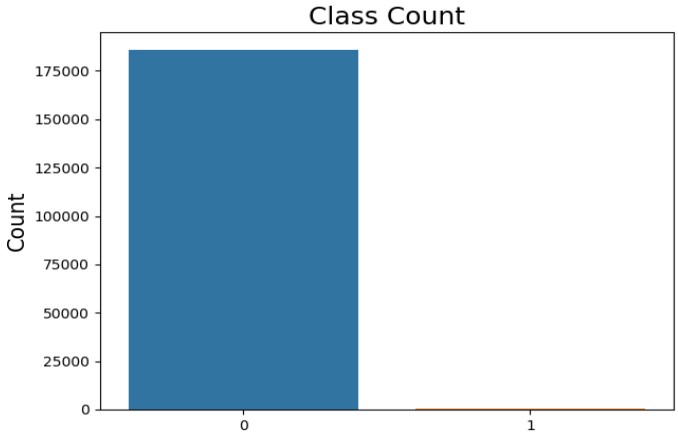


Fig 3. Class Count

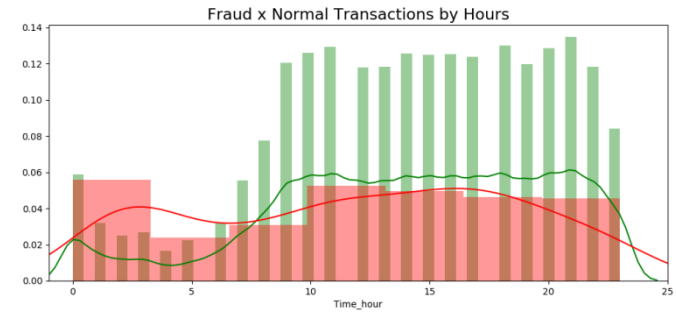


Fig 4. Fraud and Normal Transactions by Hours

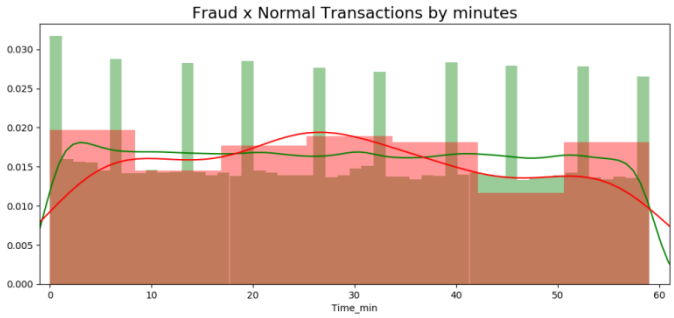


Fig 5. Fraud and Normal Transactions by Minutes

ISSN (Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 4, April 2019

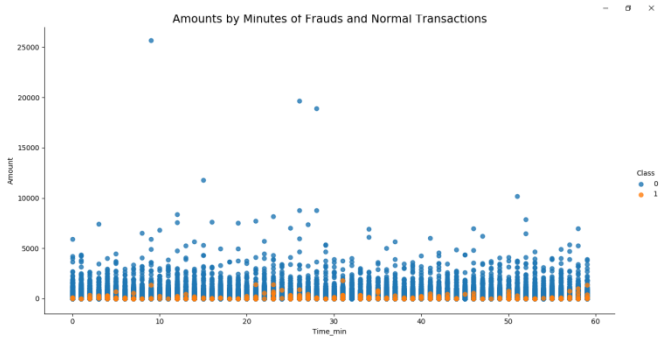


Fig 6. Amount by Minutes of Fraud and Normal Transactions

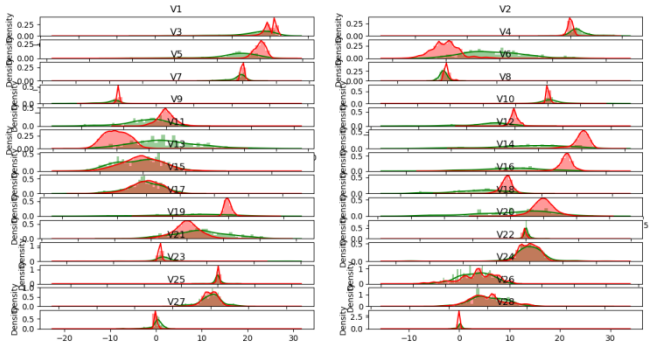


Fig 7. Fraud and Non-Fraud Detection of Each Columns

VI. CONCLUSION

In earlier stage fraudulent cases has very less because everyone have direct cash withdrawal from bank or if any urgency they go directly .but today fraudulent cases reported day by day. Although random forest obtains good results on small set data, there are still some problems such as imbalanced data. But data should be balanced by using smote technique .the accuracy level compared to other algorithm it gives more. and also choose best feature extraction and pre-processing technique to enhance the algorithm performance. From this project we detect the fraudulent cases taken in society.

REFERENCES

1. Ishu Trivedi, Monika, Mrigya Mridushi "Credit Card Fraud Detection, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, Issue 1, January 2016 pp. 39-42.
2. S.Vimala, K.C.Sharmili, Survey Paper for Credit Card Fraud Detection Using Data Mining Techniques, International Journal of Innovative Research in Science, Engineering and Technology, Vol. 6, Special Issue 11, September 2017, pp. 357-364
3. Linda Delamaire (UK), Hussein Abdou (UK), John Pointon (UK), Credit card fraud and detection techniques: a review, Banks and Bank Systems, Volume 4, Issue 2, 2009

ISSN (Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 4, April 2019

4. Andrea Dal Pozzolo, Giacomo Boracchi, Olivier Caelen, Cesare Alippi, and Gianluca Bontempi, "Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy", IEEE Transactions on Neural Networks and Learning Systems, Volume 29 Issue 8, Aug. 2018
5. E. Aleskerov, B. Freisleben, and B. Rao, "CARDWATCH: A neural network based database mining system for credit card fraud detection," in Proc. IEEE/IAFE Computat. Intell. Financial Eng., Mar. 1997, pp. 220–226.
6. C. Alippi, G. Boracchi, and M. Roveri, "A just-in-time adaptive classification system based on the intersection of confidence intervals rule," Neural Netw., vol. 24, no. 8, pp. 791–800, 2011
7. C. Alippi, G. Boracchi, and M. Roveri, "Hierarchical change-detection tests," IEEE Trans. Neural Netw. Learn. Syst., vol. 28, no. 2, pp. 246–258, Feb. 2016
8. A. C. Bahnsen, D. Aouada, and B. Ottersten, "Example-dependent cost-sensitive decision trees," Expert Syst. Appl., vol. 42, no. 19, pp. 6609–6619, 2015.
9. S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit card fraud: A comparative study," Decision Support Syst., vol. 50, no. 3, pp. 602–613, 2011.
10. R. Brause, T. Langsdorf, and M. Hepp, "Neural data mining for credit card fraud detection," in Proc. Tools Artif. Intell., Jul. 1999, pp. 103–106.
11. J. Gama, P. Medas, G. Castillo, and P. Rodrigues, "Learning with drift detection," in Advances in Artificial Intelligence. Berlin, Germany: Springer, 2004, pp. 286–295.
12. H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, Sep. 2009
13. G. Krempl and V. Hofer, "Classification in presence of drift and latency," in Proc. 11th Data Mining Workshops, Dec. 2011, pp. 596–603
14. K. Nishida and K. Yamauchi, "Detecting concept drift using statistical testing," in Discovery Science. Berlin, Germany: Springer, 2007, pp. 264–269.
15. B. Settles, "Active learning literature survey," Univ. Wisconsin, Madison, vol. 52, nos. 55–66, p. 11, 2010.
16. J. Demšar, "Statistical comparisons of classifiers over multiple data sets," J. Mach. Learn. Res., vol. 7, pp. 1–30, Jan. 2006.
17. A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi, and G. Bontempi, "Credit card fraud detection and concept-drift adaptation with delayed supervised information," in Proc. Int. Joint Conf. Neural Netw., 2015, pp. 1–8.