Credit Card Fraud Detection

M.Madhavi

Associate Professor
Department of Computer Science and Engineering
Anurag Group Of Institutions, Venkatapur, Telangana, India

G.Sumanth
Department of Computer Science and Engineering
Anurag Group Of Institutions, Venkatapur, Telangana, India

R.Hemanth
Department of Computer Science and Engineering
Anurag Group Of Institutions, Venkatapur, Telangana, India

G.Enoch
Department of Computer Science and Engineering
Anurag Group Of Institutions, Venkatapur, Telangana, India

Abstract- The project is mainly focused on credit card fraud detection in real world. A phenomenal growth in the number of credit card transactions, has recently led to a considerable rise in fraudulent activities.Many machine learning algorithms can be used to analyze the credit card fraud, the paper is focused mainly on the Random Forest algorithm because of its advantages like higher dimensionality and accuracy. It is capable to solve both classification and regression issues.

Keywords – Credit card, Fraud Detection, Accuracy, Random Forest, Dataset.

I. Introduction

Financial fraud is increases in modern communication world within seconds. Credit Card Fraud is one of the biggest threats to business establishments today. However, to combat the fraud effectively, it is important to first understand the mechanisms of executing a fraud. Credit card fraudsters employ a large number of ways to commit fraud. In simple terms, Credit Card Fraud is defined as “when an individual uses another individual’s credit card for personal reason while the owner of the card and the card issuer are not aware of the fact that the card is being used”. Card fraud begins either with the theft of the physical card or with the important data associated with the account, including the card account number or other information that necessarily be available to a merchant during a permissible transaction. Machine learning approach is based on algorithm performance, so here we use much accurate algorithm Random Forest, this is the best algorithm for classification. This 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.

II. Proposed System

2.1PROPOSED ALGORITHM UTILISED:

RANDOM FOREST ALGORITHM –

Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. Random forest is a tree-based algorithm which involves building several trees and combining with the output to improve generalization ability of the model. This method of combining trees is known as an ensemble method.Ensemble is nothing but a combination of weak learners (individual trees) to produce a strong learner. Random Forest can be used to solve regression and classification problems. In regression problems, the dependent variable is continuous. In classification problems, the dependent variable is categorical.

Bagging Algorithm is used to create random samples. Data set D1 is given for n rows and m columns and new data set D2 is created for sampling n cases at random with replacement from the original data. From dataset D1,1/3rd of rows are left out and is known as Out of Bag samples. Then, new dataset D2 is trained to this models and Out of Bag samples is used to determine unbiased estimate of the error. Out of m columns, M << m columns are selected at each node in the data set. The M columns are selected at random. Usually, the default choice of M, is m/3 for regression tree and M is sqrt(m) for classification tree. Unlike a tree, no pruning takes place in random forest i.e., each tree is grown fully.

ADVANTAGES OF PROPOSED ALGORITHM –

Pros of using random forest for classification and regression.

1. The random forest algorithm is not biased, since, there are multiple trees and each tree is trained on a subset of data. Basically, the random forest algorithm relies on the power of "the crowd"; therefore, the overall biasedness of the algorithm is reduced.

2. This algorithm is very stable. Even if a new data point is introduced in the dataset the overall algorithm is not affected much since new data may impact one tree, but it is very hard for it to impact all the trees.

3. The random forest algorithm works well when you have both categorical and numerical features. The random forest algorithm also works well when data has missing values or it has not been scaled well.

2.2 ARCHITECTURE OF PROPOSED SYSTEM

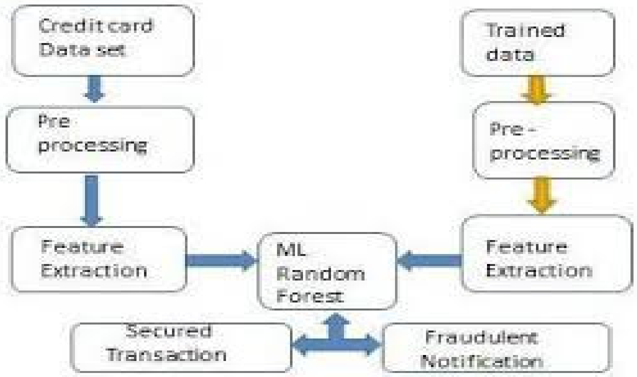


Figure 1. Workflow of credit card fraud detection

III. REQUIREMENTS SPECIFICATIONS

3.1 Technologies Requirements

PYTHON:

Python is an interpreted, high-level and general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Python is dynamically-typed and garbage-collected. Most Python implementations (including CPython) include a read–eval–print loop (REPL), permitting them to function as a command line interpreter for which the user enters statements sequentially and receives results immediately. Other shells, including IDLE and IPython, add further abilities such as improved auto-completion, session state retention and syntax highlighting.

MACHINE LEARNING:

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. Resurging interest in machine learning is due to the same factors that have made data mining and Bayesian analysis more popular than ever. Things like growing volumes and varieties of available data, computational processing that is cheaper and more powerful, and affordable data storage. All of these things mean it's possible to quickly and automatically produce models that can analyze bigger, more complex data and deliver faster, more accurate results – even on a very large scale. Because of new computing technologies, machine learning today is not like machine learning of the past. It was born from pattern recognition and the theory that computers can learn without being programmed to perform specific tasks; researchers interested in artificial intelligence wanted to see if computers could learn from data.

3.2 Software Requirements

a. PYCHARM
b. KAGGLE
c. LIBRARIES: NUMPY, PANDAS, MATPLOTLIB

IV. SYSTEM MODULES

4.1 MODULES DESCRIPTION

4.1.1 MODULE 1: DATA REVIEW

The dataset is the Kaggle credit card fraud detection data set. It contains two-day transactions made by European cardholders. The dataset contains 492 frauds out of 284,809 transactions. Thus, it is highly unbalanced, with the positive(frauds) accounting for only 0.17%.

Looking at the data shown in the video, you may find it only contains numerical variables. Features V1, V2, … V28 are the principal components obtained with PCA transformation. The only features which have not been transformed are ‘Time’ and ‘Amount’. ‘Time’ is the seconds elapsed between each transaction and the first. ‘Amount’ is the transaction amount.

‘class’ is the response variable with 1 as fraud and 0 otherwise.

4.1.2 MODULE 2:DATA PREPROCESSING

Data Preparation is the process of collecting, cleaning, and consolidating data into one file or data table, primarily for use in analysis. Data preparation is the act of manipulating (or pre-processing) raw data (which may come from disparate data sources) into a form that can readily and accurately be analyzed, e.g. for business purposes. Data preparation is the first step in data analytics projects and can include many discrete tasks such as loading data or data ingestion, data fusion, data cleaning, data augmentation, and data delivery.

4.1.3 MODULE 3:CLASSIFICATION

Data classification is broadly defined as the process of organizing data by relevant categories so that it may be used and protected more efficiently. On a basic level, the classification process makes data easier to locate and retrieve. Data classification is of particular importance when it comes to risk management, compliance, and data security. Data classification involves tagging data to make it easily searchable and trackable. It also eliminates multiple duplications of data, which can reduce storage and backup costs while speeding up the search process.

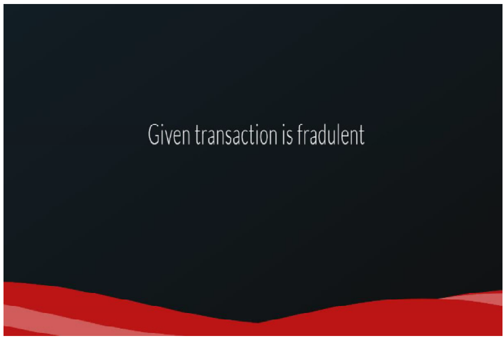
V. APPENDICES

5.1 TEST CASES:

S.NO	Test case	Time, amount, transaction method, transaction id, location, type of card, bank	Result
1.	Trying to predict the fraud or normal	values	fraud
2.	Trying to predict fraud or normal	values	Not fraud
3.	Trying to predict fraud or normal	values	fraud

4	Trying to predict fraud or normal	values	fraud

5.2 SCREENSHOTS OF RESULT FOR THE PROJECT:


Given transaction is fradulent

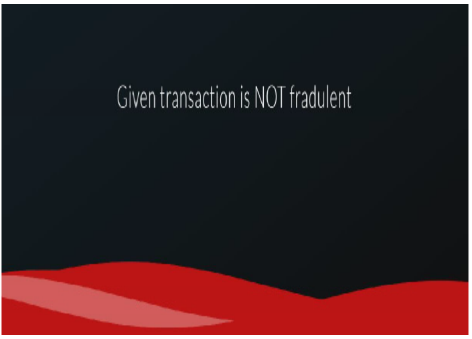

Given transaction is NOT fradulent

Figure 2. Fraud Transaction

Figure 3. Valid Transaction

IV.CONCLUSION

Credit card fraud is without a doubt an act of criminal dishonesty. we find how machine learning can be applied to methods of fraud along with their detection methods and get better results in fraud detection along with the algorithm, 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, it's 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. We got highest accuracy for Random Forest algorithm. Random Forest performing well in handling huge amount of highly imbalanced datasets in minimum amount of time. With the accuracy in the end results it shows significant growth in detecting credit card fraud transactions when compared to the algorithms like decision tree, support vector machines and logistic regression etc., because they are not performed well in handling imbalanced data sets. Random forest solves the overfitting issue by using bunch of decision trees. Random forest considered as the solution for Imbalanced classification. Random forest plays a significant role in the work-cycle of major companies because of its transparency, accuracy and its usage with different resampling techniques like SMOTE, Tomek links removal, Random under sampling, Random Over sampling. By using Data science, Artificial Intelligence and Machine learning financial institutions get benefit by saving millions of dollars every day by avoiding fraud transactions in every possible way.

REFERENCES

[1]. “Credit Card Fraud Detection Based on Transaction Behaviour -by John Richard D. Kho, Larry A. Vea” published by Proc. of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 2017

[2]. CLIFTON PHUA, VINCENT LEE1, KATE SMITH1 & ROSS GAYLER2 “A Comprehensive Survey of Data Mining-based Fraud Detection Research” published by School of Business Systems, Faculty of Information Technology, Monash University, Wellington Road, Clayton, Victoria 3800, Australia

[3]. “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) Volume 3 Issue 3, March 2014

[4]. “Research on Credit Card Fraud Detection Model Based on Distance Sum – by Wen-Fang YU and Na Wang” published by 2009 International Joint Conference on Artificial Intelligence

[5]. “Credit Card Fraud Detection through Parenclitic Network Analysis-By MassimilianoZanin, Miguel Romance, ReginoCriado, and SantiagoMoral” published by Hindawi Complexity Volume 2018, Article ID 5764370, 9 pages

[6]. “Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy” published by IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 29, NO. 8, AUGUST 2018

[7]. “Credit Card Fraud Detection” by IshuTrivedi, Monika, Mrigya, Mridushi published by International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 1, January 2016

[8]. Credit card fraud detection system using smote technique and whale optimization (sahayaskila.V,D.KayvaMonisha,Aishwaraya ,Sikhakolli)

[9]. Credit card fraud GitHub (https://github.com/sharmaroshan/Credit-Card-Fraud- Detection)

[10]. Quah, J. T. S., and Sriganesh, M. (2008). Real-time credit card fraud detection using computational intelligence. Expert Systems with Applications, 35(4), 1721-1732.
[11]. Deep fake analysis by Ananta prabhu G cybercrimes expert Mangalore police department.

[12]. L. Daly, "Identity theft credit card fraud statistics for 2020," 13 April 2020. [Online]. Available: Https://www.fool.com/the-ascent/research/identity-theft-credit-card-fraud-statistics/.

[13]. D. Clark, "Distribution of Fraud losses on UK-issued debit and credit cards in 2019, by type," statista.com, 24 November 2020. [Online]. Available: https://www.statista.com/statistics/286282/united-kingdom-uk-fraud-losses-on-plastic-cards-by-type/.redit Card Fraud and Cyber Crime: Cyber Crime: The Intersection | SQN Banking Systems