Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Credit Card Fraud Detection Using Data Science Techniques

A.Banupriya[1], Divya.R[2], M.Vinodhini[3]
[1]*Assistant Professor Department of Information Technology, CSI College of Engineering, Ketti, Tamil Nadu, India*
[2,3]*UG Scholar, Department of Information Technology, CSI College of Engineering, Ketti, Tamil Nadu, India*
vishmithaammu@gmail.com

----------**********----------

Abstract:
Credit card fraud is a phenomenon which is having a significant impact on a social life. Credit card fraud detection is becoming apparent research area which is obtaining interest but involved some challenges due to the limited amount of resources (i.e., datasets) accessible. In this paper we propose, credit card fraud detection model that uses data science techniques like Random Forest algorithm (machine learning) and Convolutional Neural Network algorithm (deep learning) for analysis. We investigate and compare the accuracy of the above mentioned algorithms (i.e., RFA, CNN) and create interface to predict whether the transaction is genuine or fraudulent.Credit card fraud increases as ecommerce that becomes more prevalent.

Keywords **—Random Forest, Convolutional Neural Network, ecommerce**

----------**********----------

I. INTRODUCTION

Credit card transaction datasets are rarely available, highly imbalanced and skewed. The most favorable feature (variables) selection for the models, worthy measure is most important part of data mining to evaluate performance of techniques in credit card fraud data. A number of challenges are related with credit card detection, that is, fraudulent behavior profile is dynamic, that is fraudulent transactions tend to look like legalized ones. Performance is greatly affected by type of sampling approaches used, selection of variables and detection techniques are used. In the end of this paper, conclusions about results of classifier evaluative testing are made and collated. From the experiments, the result that has been concluded is that the Logistic regression has a accuracy of 97.7%, while Support Vector Machine 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%.

1.1 Deep Learning

Deep learning is actually a subdivision of machine learning. It practically is machine learning and functions in the same

way but it has dissimilar capabilities. The overall difference between the Deep Learning and Machine Learning is, machine learning models become well progressively but the model still needs some guidance. If a machine learning model returns a wrong prediction then the programmer needs to fix that problem clearly but in the case of deep learning, the model does it by him. In particular Automatic car driving system in deep learning. Deep learning is a part of machine learning with an algorithm inspired by the structure and function of the brain, which is called an **artificial neural network**. Deep learning is suited over a range of fields such as computer vision, speech recognition, natural language processing, etc.

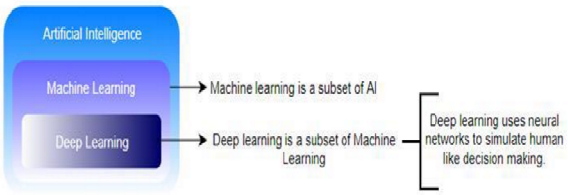

Fig 1.1 Artificial Intelligence

AI stands for Artificial Intelligence. It is a technique which enables machines to mimic human behavior. Machine Learning is a subset of AI which uses statistical methods to enable machines to improve with experiences. Deep learning is a part of Machine learning, which makes the computation of multi-layer neural networks feasible. It

takes advantage of neural networks to simulate human-like decision making.

Deep Learning, as a branch of Machine Learning, employs algorithms to process data and imitate the thinking process, or to develop abstractions. Deep Learning (DL) uses layers of algorithms to process data, acknowledge human speech, and visually realize objects.

1.2 Importance of DL

Deep learning is a strong tool to make prediction a working result. Deep learning stand out in pattern discovery (unsupervised learning) and experimental prediction. Big data is the fuel for deep learning. When both are combined, an organization can secure unprecedented results in term of productivity, sales, management, and innovation. Deep learning can outperform traditional method. For example, deep learning algorithms are 41% more accurate than machine learning algorithm in image classification, 27% more accurate in facial recognition and 25% in voice recognition.

1.3 Limitations of DL

Data labeling

Most modern AI models are trained through "supervised learning." It means that humans must label and classify the underlying data, which can be a significant and error-prone chore. For example, companies developing self-drivingcar technologies are hiring hundreds of people to manually explain hours of video feeds from prototype vehicles to help train these systems.

Obtain huge training datasets

It has been shown that simple deep learning techniques like CNN can, in some cases, follow the knowledge of experts in medicine and other fields. The current sign of machine learning, however, requires training data sets that are not only labeled but also relatively broad and universal. Deep-learning methods need thousands of survey for models to become reasonably good at classification tasks and, in certain cases, millions for them to perform at the level of humans. Without revelation, deep learning is famous in giant tech companies; they are using big data to gather petabytes of data. It allows them to create an magnificent and highly accurate deep learning model.

1.4 DL Process

A deep neural network gives state-of-the-art accuracy in countless tasks, from object detection to speech recognition. They can learn spontaneously, without pre-established knowledge specifically coded by the programmers.

Fig 1.4 DL Process

To hold the idea of deep learning, visualize a family, with a baby and parents. The child points objects with his little finger and always says the word 'cat.' As its parents are worried about his education, they keep telling him 'Yes, that is a cat' or 'No, that is not a cat.' The infant persists in pointing objects but becomes more accurate with 'cats.' The little kid, deep down, does not know why he can say it is a cat or not. He has just learned how to order complex features coming up with a cat by looking at the pet overall and continue to focus on features such as the tail feathers or the nose before to make up his mind. A neural network works quite the same. Each layer mean a greater level of knowledge, i.e., the ranking of knowledge. A neural network with four layers will learn many complex feature than with two layers.

The learning occurs in two phases.

1. The first phase consists of applying a nonlinear transformation of the input and creates a statistical model as output.
2. The second phase aims at improving the model with a mathematical method known as derivative. The neural network repeats these two phases hundreds to thousands of time until it has reached a tolerable level of accuracy. The repeat of this two-phase is called iteration.

1.5 Machine Learning Vs Deep Learning

Deep learning is a detailed form of machine learning. A machine learning workflow starts with related features being manually obtained from images. The features are then used to create a model that categorizes the objects in the image. With a deep learning workflow, relevant features are automatically derived from images. In addition, deep learning accomplish “end-to-end learning” – where a network is given primary data and a task to perform, such as classification, and it learns how to do this automatically.

Another key differentiation is deep learning algorithms scale with data, whereas shallow learning coincide. Shallow learning refers to machine learning methods that highland at a certain level of performance when you add more examples and training data to the network.

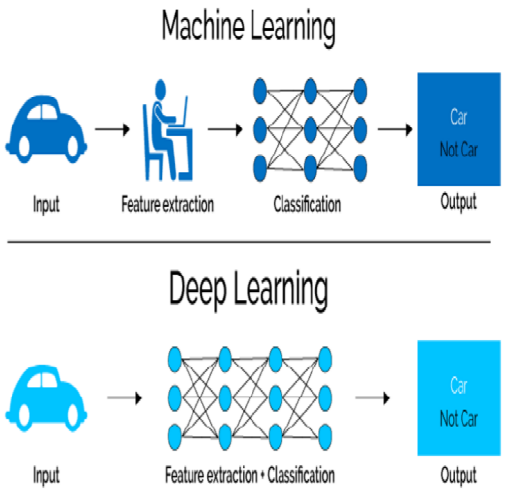

Fig 1.5 ML Vs DL 1.5.1

How DL Works?

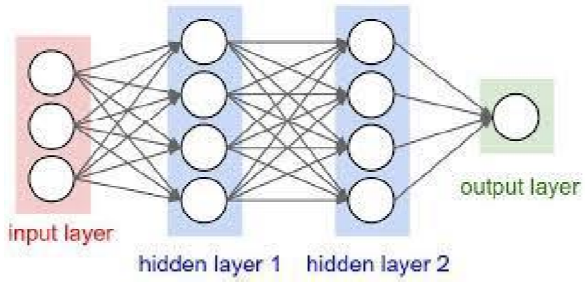

Fig 1.5.1 Working of DL

Deep learning has developed hand-in-hand with the digital epoch, which has brought about an ignition of data in all forms and from every sector of the world. This data, known quietly as big data, is obtained from sources like social media, internet search engines, e-commerce platforms, and online cinemas, among others. This enormous amount of data is readily accessible and can be shared through applications like cloud computing.

Although, the data, which normally is unstructured, is so wide that it could take decades for humans to understand it and obtain applicable information. Companies realize the unbelievable potential that can result from untangle this wealth of information and are increasingly adapting to AI systems for automated support.

Many deep learning methods use **neural network** architectures, which is why deep learning models are frequently referred to as **deep neural networks**. The term "deep" generally refers to the number of hidden layers in the neural network. Conventional neural networks only contain 2-3 hidden layers, while deep networks can have as many as 150. Deep learning models are trained by using large sets of categorized data and neural network architectures that learn features straight away from the data without the need for manual feature extraction.

1.5.2 Applications of DL:

- Image recognition
- Natural language processing
- Speech recognition
- Robots and self-driving cars

1.6 Training DL Model

How to Create and Train Deep Learning Models? The three most common ways people use deep learning to perform object classification is:

Training from Scratch

To train a deep network from scratch, we collect a very large categorized data set and design a network architecture that will learn the features and model. This is good for applications that will have a large number of output categories. This is a rare approach because with the large amount of data and rate of learning, these networks generally take days or weeks to train.

Transfer Learning

Most deep learning applications use the transfer learning approach, a approach that involves fine-tuning a pretrained model. You start with a subsist network, such as AlexNet or GoogLeNet, and feed in new data containing previously unspecified classes. After making some twist to the network, you can now perform a new task, such as classifyng only dogs or cats instead of 1000 different objects. This also has the advantage of requiring much less data (processing thousands of images, rather than millions), so computation time drops to minutes or hours. Transfer learning needs an interface to the internals of the pre-existing network, so it can be accurately modified and enhanced for the new task. MATLAB has tools and functions designed to help you do transfer learning.

Feature Extraction

Moderately less common, more specialized approach to deep learning is to use the network as a **feature extractor**. Since all the layers are assigned with learning certain features from images, we can pull these features out of the network at any time during the training process. These

features can then be used as input to a machine learning model such as support vector machines (SVM).

Accelerating Deep Learning Models with GPUs
Training a deep learning model can take a long period, from days to weeks. Using GPU acceleration can speed up the process notably. Using MATLAB with a GPU lessen the time required to train a network and can cut the training time for an image classification problem from days down to hours. In training deep learning models, MATLAB uses GPUs without requiring you to understand how to program GPUs clesrly.

Types of Deep Learning:
There are three types of deep learning networks. They are

1. Feed forward neural networks
2. Recurrent neural networks (RNN)
3. Convolutional neural networks (CNN)

Feed-Forward Neural Networks
The simplest type of artificial neural network. With this kind of architecture, the information flows in only one direction, forward. This means that, the information's flows starts at the input layer, pass to the "hidden" layers, and end at the output layer. The network does not have a loop. Information stops at the output layers.

Recurrent Neural Networks (RNNs)
Recurrent Neural Network is a multi-layered neural network that can store information in condition nodes, allowing it to learn data series and output a number or another series. In easy words it an artificial neural networks which connections between neurons include loops. RNNs are well suited for processing series of inputs.

1.7 Convolutional Neural Networks (CNN)
Convolutional Neural Networks is a multi-layered neural network with a distinctive architecture designed to extract progressively complex features of the data at each layer to determine the output. CNN's are well suited for perceptual tasks.

CNN is mainly used when there is an unstructured data set (e.g., images) and the professionals need to extract information from it. For example, if the task is to predict an image caption: The CNN acquires an image of a cat, this image, in computer term, is a collection of the pixel. Basically, one layer for the greyscale picture and three layers for a color picture. During the feature learning (i.e., hidden layers), the network will recognize unique features, for example, the tail feathers of the cat, the ear, etc. When the network completely learned how to realize a picture, it can provide a probability for each image it knows. The label with the highest possibility will become the prediction of the network.

II. LITERATURE SURVEY

1. BLAST-SSAHA Hybridization for Credit Card Fraud Detection (Author:-AmlanKundu, SuvasiniPanigrahi, ShamikSural, Senior Member, IEEE, and Arun K. Majumdar, Senior Member, IEEE)
2. Detecting Credit Card Fraud by Decision Trees and Support Vector Machines (Author:-Y. Sahin and E. Duman)
3. Evaluating and Emerging payment card fraud challenges and resolution," International Journal of Computer Applications, vol. (Author:-P. Richhariyaand P. K. Singh)
4. Card fraud detection using learning machines(Author:-A. E. Pasarica)
5. Learned lessons in credit card fraud detection from a practitioner perspective(Author:-A. Dal Pozzolo, O. Caelen, Y.-A. Le Borgne, S. Waterschoot, and G. Bontempi)

2.1 System Analysis

2.1.1 System Specifications

2.1.1.1 Hardware Specification
Windows 10 64 bit
RAM 4GB
2.1.1.2 Software Specification
Python v3.6.8
Python idle

2.2 Existing System:

In credit card fraud detection, Data normalization is applied before Cluster Analysis and with results obtained from the use of Cluster Analysis 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. Consequence of this paper was to find new methods for fraud detection and to increase the accuracy of results.
The data set for this paper is based on real life transactional data by a huge European company and personal details in data is kept sensitive. Accuracy of an algorithm is around 50%. Significance of this paper was to perceive an algorithm and to lessen the cost measure. The result achieved was by 23% and the algorithm they find was Bayes minimum risk.

2.2.1 Disadvantages

In this paper a new combined comparison measure that reasonably means the gains and losses due to fraud detection is proposed.

A cost delicate method which is based on Bayes minimum risk is presented using the proposed cost measure.

III. PROPOSED SYSTEM

We are applying random forest algorithm and convolutional neural network algorithm to compare the accuracy of the credit card datasets.

Random Forest (RF) is an algorithm for classification and regression. It is a collection of decision tree classifiers. Random forest has benefit over decision tree as it corrects the habit of over fitting to their training set. A subset of the training set is selected randomly so that to train each isolated tree and then a decision tree is built; each node then divide on a feature selected from a random subsequence of the full feature set.

Even for huge data sets with many features and data occurrence training is extremely fast in random forest and because each tree is trained independently of the others. The RF algorithm has been established to provide a good estimate of the generalization error and to be resistant to over fitting.

Convolutional Neural Network are neural networks that share their parameters. CNN models were developed for image classification problems, where the model studies an internal representation of a two-dimensional input, in a process referred to as feature learning. There are 3 dimensions in CNN, where the inputs will be in the form of datasets, images (i.e., 2D & 3D).

3.1 Advantages o

Reduces human work o

Less time consumption

o Better performance

3.2 Block Diagram

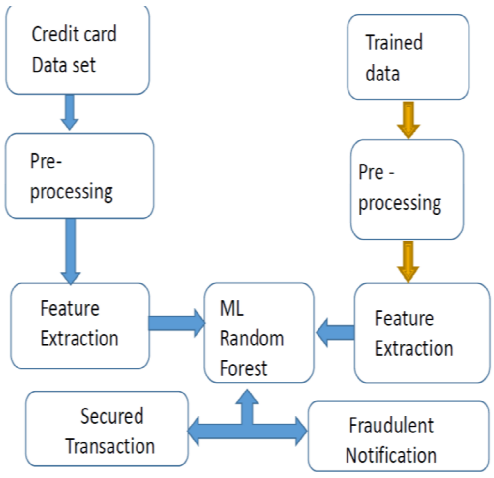

Fig .3.2 Block Diagram

3.3 Dataset:

The datasets contains transactions made by credit cards. The dataset is highly unbalanced. It contains only numerical input variables which are the result of a PCA transformation cannot provide the original features and more background information about the data.

Characteristics V1, V2, ... V28 are the principal components gained with PCA, the only features which have not been changed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds proceed between each transaction and the first transaction in the dataset. The characteristic 'Amount' is the transaction Amount, this feature can be used for example-dependent cost-sensitive learning. Characteristic 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise."

3.4. Preprocessing

Data preprocessing which largely include data cleaning, integration, transformation and reduction, and attain training sample data needed.

This is a data mining strategy that changes raw data into an understandable format.

Steps in Data Preprocessing

1. Import libraries
2. Read data
3. Checking for missing values
4. Checking for categorical data
5. Standardize the data
6. PCA transformation
7. Data splitting

3.4.1 Feature Extraction

Feature selection include reducing the computational costs, saving storage space, Facilitating model selection procedures for accurate prediction, and interpreting complex dependencies between variables. The features that are well selected not only optimize the classification accuracy but also reduce the number of required data for achieving an optimum level of performance of the learning process. Feature selection methods usually include search strategy, assessment measure, stopping criterion, and validation of the results. Search strategy is a search method used for producing a subset of candidate features for assessment. An assessment measure is applied for evaluating the quality of the subset of candidate features. Validation is the study of validity of the selected features with the real world datasets.

3.5 Random Forest Algorithm

It is the basic classifier and it establishes a large number of trees. Random Forests is a powerful prediction tool commonly used in data mining. It establishes a series of classification trees which will be used to classify a new instance. The idea used to create a classifier model is establish multiple decision trees, each of which uses a subset of attributes randomly selected from the entire original set of attributes.

Candidate split dimension along which a split may be made. Candidate split point one of the first m structure points to appear in a leaf. Candidate split a combination of a contestant break dimension and a location along that dimension to split.

They are formed by projecting each contestant split point into each candidate split dimension. Candidate children each candidate split in a leaf originate two candidate children for that leaf. They are also referred to as the left and right child of that split.

Trained Data

The quality, variety, and quantity of our training data decide the success of our machine learning models. The form and content of the training data often referred to as categorized or human categorized data or ground truth dataset is designed for to train particular ML models with an end application in perspective.

3.6 CNN

CNN or covnets are neural networks that can share their parameters. Imagine you have an image. It can be illustrated as a cuboid having its length, width and height (as image generally have red, green, and blue channels). Convolutional neural network models were established for image classification issues, where the model learns an internal representation of a 2D input, in a process referred to as feature learning. This same process can be controlled on 1D sequences of data. The model extracts features from sequences data and maps the internal features of the sequence. A one-dimensional Convolutional Neural Network is very effective for acquiring features from a fixed-length segment of the overall dataset, where it is not important where the feature is located in the segment.

3.6.1 1D CNN

Evaluation of a time series of sensor data. Evaluation of signal data over a fixed-length period, for example, an audio recording.

Natural Language Processing (NLP), although Recurrent Neural Networks which leverage Long Short Term Memory (LSTM) cells are more promising than CNN as they take into account the proximity of words to create trainable patterns. Likewise, 1D Convolutional Neural Networks are also used on audio and text data since we can also represent the sound and texts as a time series data.

IV. CONCLUSION

This paper has examined the performance of two kinds of algorithms such as Random Forest algorithm and Convolutional Neural Network algorithm. A real-life dataset on credit card transactions is used in our experiment. Although random forest obtains good results on small set data, there are still some problems such as imbalanced data. Our future work will focus on solving these problems. The algorithm of random forest itself should be improved. For example, the voting mechanism assumes that each of base classifiers has equal weight, but some of them may be more important than others.

References

1. ” Credit card Fraud Detection System using Hidden Markov Model and Adaptive Communal Detection”, International Journal of Computer Science and Information Technologies, vol 6 (2), 2015.

2. “Cost sensitive Modeling of Credit Card Fraud Using Neural Network strategy”, ICSPIS 2016, 14-15 Dec 2016, Amirkabir University of Technology Tehran, Iran.

3. Analysis on Credit Card Fraud Detection Methods”International Journal of Computer Trends and Technology (IJCTT) – volume 8 number 1– Feb 2014

