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 5, May 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 5, May 2021||

DOI:10.15680/IJIRSET.2021.1005165

Credit Card Fraud Detection Using Machine Learning

Prof. Santosh Jadhav[1], Pratik Borkar[2], Ankita Munj[3], Prathmesh Risbud[4], Sahil Sarang[5]

Associate Professor, Department of Information Technology, Finolex Academy of Management and Technology, Ratnagiri, Maharashtra, India[1]

Student, Department of Information Technology, Finolex Academy of Management and Technology, Ratnagiri, Maharashtra, India[2]

Student, Department of Information Technology, Finolex Academy of Management and Technology, Ratnagiri, Maharashtra, India[3]

Student, Department of Information Technology, Finolex Academy of Management and Technology, Ratnagiri, Maharashtra, India[4]

Student, Department of Information Technology, Finolex Academy of Management and Technology, Ratnagiri, Maharashtra, India[5]

ABSTRACT: The uses of credit cards are rising day by day as modern world is going digital use of electronic currency also increased. The benefits of credit card is you don't need to carry cash every time as it can be used online as well as offline transactions this increases popularity of the credit cards in public. As a result numbers of frauds cases are also increasing in this sector so it becomes important to detect frauds. By detecting frauds credit card companies can keep their customers money safe and this will also helpful in increasing trust of the customers on the banking system. General idea of this paper is to classify fraudulent and genuine transactions using various machine learning techniques as it can be fastest and easiest way to detect fraudulent transactions. We purpose to use Gaussian Naive Bays, Logistic Regression, Linear Discriminant Analysis algorithms from supervised learning techniques with SMOTE sampling and hyperparameter tuning for handling imbalance in dataset and Local Outlier Factor, Isolation Forest from unsupervised learning techniques. We will also perform comparative analysis it will help us to understand which algorithm performs better.

KEYWORDS: credit card, fraud detection, machine learning, data science, smote, sampling.

I. INTRODUCTION

Nowadays with the rapid increase in digital payment modes, frauds in the banking sectors are also increased. Fraud in banking is nothing but getting money by deceiving the peoples. Scammers are adapting smart methodologies for performing various frauds that's why banking systems also need to be updated with the latest fraud detection techniques. One of the most popular modes of payment is through credit cards as it is the most used payment mode frauds are also rising day by day. Credit card fraud is an illegal act of using credit card information without the knowledge of the cardholder. Based on statistical data stated in [1] in 2012, the high-risk countries facing credit card fraud threat is Ukraine has the most fraud rate with staggering 19%, which is closely followed by Indonesia at 18.3% fraud rate. After these two, Yugoslavia with a rate of 17.8% is the riskiest country. The next highest fraud rate belongs to Malaysia (5.9%), Turkey (9%), and finally the United States. So it becomes the most critical issue worldwide. Detecting credit card fraud has become more important as it is beneficial for both the parties that are costumers and the banking system also. It will not only save customers money but also help to increase the trust of customers in the banking system.

The process of applying a computer-based data methodology to discover knowledge from data is called data mining so we can detect frauds in the system by using machine learning as frauds generally occur in a pattern. Our research is

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

||Volume 10, Issue 5, May 2021||

DOI:10.15680/IJIRSET.2021.1005165

about classifying fraudulent and non-fraudulent transactions using various machine learning techniques and study them comparatively for better results. The main challenges in this project are an enormous amount of data, the model build must be fast enough to respond, highly imbalanced dataset i.e. most of the transactions are non-fraudulent and data availability. As credit card transactions are considered as the most confidential information of the customers so baking authorities will not be ready to share this information with anyone. We collected a dataset of credit card transactions from the Kaggle website which contains 2,84,807 transactions these transactions were made by European cardholders in September 2013. The dataset contains only numerical inputs which are PCA transformed as it is confidential information so they didn't provide original components. These are transactions made by European cardholders in two days. Kaggle is the world's largest data science community with powerful tools and resources to help you achieve your data science goals[2]. The imbalance in data can be handled using various sampling techniques. We are using SMOTE oversampling along with a good amount of hyperparameter tuning because it is beneficial here as we are interested in the minority class. We have also used various data visualization techniques to draw graphs and tables. We used Gaussian Naive Bays, Logistic Regression,

Linear Discriminant Analysis from supervised learning techniques and Local Outlier Factor, Isolation Forest from unsupervised learning techniques and also perform a comparative analysis of the result by using evolution metrics such as precision, recall, F1 score and determine which model perform better..

II. LITERATURE SURVEY

Fraud act as the unlawful activity intended to result in personal benefit. It is an act that is against the law, with an aim to get unauthorized financial benefit. A survey conducted by Clifton Phua and his associates has stated that techniques employed during this domain include data processing applications, automated fraud detection, adversarial detection. In other paper, Suman, Research Scholar, GJUS&T at Hisar HCE presented techniques on Supervised and Unsupervised Learning for credit card fraud detection. Even though these methods and algorithms fetched an unexpected success in some areas, they did not provide a permanent and consistent solution to fraud detection. In a similar research domain, they used Outlier mining, Distance sum algorithms, and Outlier detection mining to accurately predict fraudulent transactions in an emulation experiment of MasterCard transaction data set of 1 certain full-service bank. Outlier mining may be a field of knowledge mining that is essentially utilized in monetary and internet fields. It deals with detecting objects that are detached from most systems i.e. the transactions that aren't genuine. They have taken attributes of customer's behavior and supported the worth of these attributes they've calculated the distance between the observed value of that attribute and its predetermined value. Unconventional techniques like hybrid data mining/complex network classification algorithm are in a position to perceive illegal instances in an actual card transaction data set, supported network reconstruction algorithm that allows creating representations of the deviation of 1 instance from a reference group have proved efficient typically on medium-sized online transaction. There have also been efforts to progress from a totally new aspect. Attempts are made to enhance the alert-feedback interaction just in case of fraudulent transactions. In case of fraudulent transactions, the authorized system would be alerted and feedback would be sent to deny the continued transaction. In the Artificial Genetic Algorithm, one of the approaches that shed new light during this domain countered fraud from a special direction. It finds out the accurate fraudulent transactions and minimizes the number of false alerts.

III. METHODOLOGY

We intended to classify the dataset transactions as fraudulent and non-fraudulent. In this paper, we used five different algorithms to detect fraud. We have taken the dataset from the Kaggle website and we are using SMOTE oversampling for handling the imbalance of data. We compared different performance metrics to determine which algorithm performed better which can be adopted to identify fraudulent activities. We will see modules included in this project.

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

||Volume 10, Issue 5, May 2021||

DOI:10.15680/IJIRSET.2021.1005165

A. Modules -

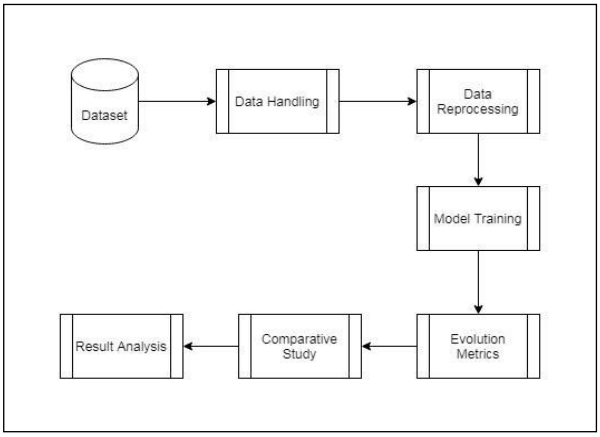

fig.1: Work Flow Diagram

In the above figure, we can clearly see all of the modules involved in this project. We will discuss them in brief.

a. Dataset - Collecting credit card data is a very tedious process as it is very secured data of the cardholders as a result bank authorities are also not ready to public them. So, the Kaggle community helpful in providing such datasets for training and testing purposes.

b. Data Handling - Data handling gives us more idea about the dataset. In this, we visualize data graphically and understand the structure of data by checking additional details using various tools.

c. Data Preprocessing - In this step various data preprocessing techniques are implemented such as scaling data by using standard scalar, separating data in training and testing sets, and using sampling for handling imbalance. Handling imbalance is a very much important step in this project because of the high imbalance structure dataset that is the dataset contains very less amount of fraud cases and they are most important.

d. Model Training - In this using various machine learning techniques we train our model. Here, five different algorithms are used namely Gaussian Naive Bays, Logistic Regression, Linear Discriminant Analysis, Local Outlier Factor, and Isolation forest.

e. Comparative Study - The comparative study is important to detect the best model for fraud detection. We can use various evolution metrics such as precision, recall, f1 score for comparing the performance of all trained models.

B. Tools and Techniques -

i. Machine learning techniques **-**

Here we are using five different algorithms form machine learning. They are briefly explained below.

Ÿ Gaussian Naive Bays - Naive Bayes are a group of supervised machine learning classification algorithms based on the Bayes theorem. It is a simple classification technique, but has high functionality. Naive Bayes Classifiers are based on the Bayes Theorem. In a supervised learning situation, Naive Bayes Classifiers are trained very efficiently.. Naive Bayes Classifiers have simple design and implementation and they can applied to many real life situations.

Ÿ Logistic Regression - It is another supervised learning algorithm which is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.

Ÿ Linear Discriminant Analysis -It is a generalization of Fisher's linear discriminant, a method used in statistics and other fields, to find a linear combination of features that characterizes or separates two or more classes of objects or

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

||Volume 10, Issue 5, May 2021||

DOI:10.15680/IJIRSET.2021.1005165

events.It is generally used as dimensionality reduction technique which is commonly for the supervised classification problems. It is used for modeling differences in groups i.e. separating two or more classes.

Ÿ Local Outlier Factor -Local outlier factor (LOF) is an algorithm used for Unsupervised outlier detection. It produces an anomaly score that represents data points which are outliers in the data set. It does this by measuring the local density deviation of a given data point with respect to the data points near it.

Ÿ Isolation Forest -Isolation forest is an unsupervised learning algorithm for anomaly detection that works on the principle of isolating anomalies. In order to isolate a data point, the algorithm recursively generates partitions on the sample by randomly selecting an attribute and then randomly selecting a split value for the attribute, between the minimum and maximum values allowed for that attribute.

ii. Data –

The dataset is collected from kaggle which is the world's largest data science community with powerful tools and resources. The dataset contains transactions made by credit cards in September 2013 by European cardholders contains 284,807 transactions out of which 492 transactions are frauds i.e. dataset is highly unbalanced. It contains only numerical input variables which are the result of a PCA transformation due to confidentiality issues original features cant be provided. Only components that are not PCA transformed are time and amount.

iii. Sampling technique –

As we can see our dataset is highly imbalance so there is need of sampling techniques as undersampling includes risk of important data loss we chosen to use oversampling method SMOTE (Synthetic Minority Oversampling Technique). It aims to balance class distribution by randomly increasing minority class examples by replicating them.SMOTE synthesizes new minority instances between existing minority instances. It generates the virtual training records by linear interpolation for the minority class.

iv. Tools –

This proposed model is implemented in Python using interactive data science environment Jupyter Notebook. Numpy and Pandas are used for simpler tasks such as data storage and transformation. For data analysis and visualization Matplotlib is used. Seaborn is used for statistical data visualization and algorithms and preprocessing we used are from Scikit-learnlibrary.

IV. EXPERIMENTAL RESULTS

A. Evolution Metrics -

As our project is about detecting fraudulent transactions from dataset and dataset is highly imbalanced as like many other projects accuracy is not important metrics here as accuracy is assuming that there are equal quantity of examples from each class. So, accuracy is not correct measure of performance efficiency in our case to overcome this problem we chosen following standards :

- Precision
- Recall
- F1 Score

To understand this performance measures we first need to understand confusion matrix because they all are dependent on the confusion matrix. So, we draw 2 X 2 sample confusion matrix.

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

||Volume 10, Issue 5, May 2021||

DOI:10.15680/IJIRSET.2021.1005165

	Predicted Class		
Actual Class		Negative	Positive
	Negative	True Negative (TN)	False Positive (FP)
	Positive	False Negative (FN)	True Positive (TP)

Table 1: Confusion Matrix

- True positives (TP) - These are cases in which we predicted positive and it is true that means value of both actual class and predicted class are positive.
- True negatives (TN) - In this we predicted negative, and it is true which means that value of both actual class and predicted class are negative.
- False positives (FP) - We actually predicted positive but it is false that means value in actual class is negative and value in predicted class is positive it is also known as a Type I error.
- False negatives (FN) - We actually predicted negative but it is false that means value in actual class is positive and value in predicted class is negative it i also known as a Type II error.

Precision - Out of all the positive classes we have predicted correctly, how many are actually positive.

Recall - Out of all the positive classes, how much we predicted correctly. It should be high as possible.

$$R = \frac{TP}{TP + FN}$$

F1 Score: It is the weighted average of Precision and Recall. Therefore this score takes both false negatives and false positives into account. It is harmonic mean of precision and recall.

$$P = \frac{TP}{TP + FP}$$

Support - Support is the number of samples of the true response that lie ineach class of target values.

B. Result Analysis -

We analyze the dataset to classify the transactions fraudulent and non fraudulent. Now we preform analysis on all of the final outputs and comparatively study all the algorithms and techniques.

In supervised learning we implemented three different algorithms namely Gaussian Naive Bays, Logistic Regression, Linear Discriminant Analysis. .For supervised learning we divided dataset in 80% and 20% i.e. 80% data is for training and 20% data is for testing. So, our testing set contains total 56,962 entries out of which 56,864 transactions are genuine and 98 transactions ar $F1\ Score\ = 2 \times \frac{Recall \times Precision}{Recall + Precision}$ Below given table includes all of supervised algorithms that we implemente)TE sampling.

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

||Volume 10, Issue 5, May 2021||

DOI:10.15680/IJIRSET.2021.1005165

	Sr. No.	Algorithm Name	Precision	Recall	F1 Score
No Sampling	1	Gaussian Naive Bays	0.06	0.82	0.11
	2	Logistic Regression	0.86	0.58	0.70
	3	Linear Discriminant Analysis	0.87	0.74	0.80
SMOTE Sampling	1	Gaussian Naive Bays	0.06	0.87	0.11
	2	Logistic Regression	0.06	0.92	0.11
	3	Linear Discriminant Analysis	0.10	0.84	0.19
Best SMOTE Ratio	1	Gaussian Naive Bays	0.06	0.85	0.12
	2	Logistic Regression	0.80	0.81	0.80
	3	Linear Discriminant Analysis	0.88	0.73	0.80

Table 2: Supervised learning technique outputs

From above table we can clearly see that without sampling LDA performed better with 0.80 F1 score and Logistic Regression also performed better with 0.70 F1 score but main problem here is model is biased towards majority class and can neglect minority class. As undersampling can lost important data from the dataset therefore we are using SMOTE oversampling. After using SMOTE oversampling F1 score is much suffered because problem here is SMOTE created too many entries that have too much details and produces noise in the data. To overcome this problem we choose best sampling strategy by performing hyerparameter tuning and we got best F1 scores. From above table we can clearly state that Logistic Regression and Linear Discriminant Analysis performed better.

Sr. No.	Algorithm Name	Precision	Recall	F1 Score	Accuracy
1	Local Outlier Factor	0.00	0.00	0.00	0.9953
2	Isolation Forest	0.57	0.34	0.43	0.9973

Table 3: Unsupervised learning technique outputs

FOR AN UNSUPERVISED LEARNING WE HAVE TAKEN PART OF DATASET TO REDUCE PROCESSING POWER REQUIRED FOR COMPUTING. WE TAKEN FIRST 50,000 ENTRIES FROM DATASET OUT OF WHICH 49,852 TRANSACTIONS ARE LEGITIMATE WERE AS 148 TRANSACTIONS ARE FRAUDULENT. WE ARE USING SUPERVISED SENSE HERE TO DETECT PERFORMANCE OF MODEL. FROM ABOVE TALE WE CAN CLEARLY SAY THAT ISOLATION FOREST PREFORM BEST IN SUCH SMALL AMOUNT OF DATA ADDING MORE DATA WILL IMPROVE ITS EFFICIENCY. FOR BETTER UNDERSTANDING WE DRAW F1 SCORES OF ALL ALGORITHMS.

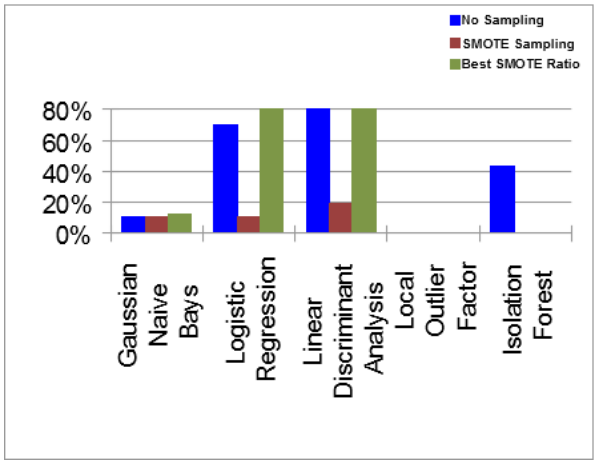

fig. 2: F1 score of various fraud detecting algorithms

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

||Volume 10, Issue 5, May 2021||

DOI:10.15680/IJIRSET.2021.1005165

V. FUTURE WORK

In this paper we tried to classify fraudulent and non-fraudulent transactions still we couldn't reach out to the goal of providing 100% accuracy for detecting frauds. We developed models which are very much close to that goal. So, there is the room for improvement. In future we can try various different algorithms and combine them for better results. We can also use good parameter tuning for getting best performance of model. Applying more preprocessing steps such as scaling (e.g. Minmaxscalar) and sampling (e.g. Borderline SMOTE, ADAYSN) can be able to improve model performance. We have seen that as size of data increases models become more accurate, hence introducing more data will also help in improving model performance but this will require official support from bank authorities.

VI. CONCLUSION

Overall objective of this project is to determine best algorithm for detecting credit card frauds. The proposed system is implemented in python coding language. In supervised learning technique Linear Discriminant Analysis performed to much better but this method generally used for dimension reduction it is not that good for use model training method so we can say that Logistic Regression also performs better. So fraud detection systems can consider Logistic Regression as a good option. On the other hand in supervised learning technique Isolation forest provides better results with such small amount of data its efficiency can be increased by providing more data. As more amount of data introduced Isolation Forest efficiency increases so the fraud detection systems can consider it as best option.

ACKNOWLEDGEMENT

The author would like to Prof. Santosh Jadhav department of IT for their support, guidance and advice for the development of this project and making it possible.

REFERENCES

[1]KhyatiChaudhary, JyotiYadav, BhawnaMallick, " A review of Fraud Detection Techniques: Credit Card", International Journal of Computer Applications Volume 45– No.1 2012.

[2]Machine Learning Group - ULB, kaggle datast for credit card fraud detection, Accessed: May, 20, 2020. [Online]Available:https://www.kaggle.com/mlg-ulb/creditcardfraud

[3]Hyder John, Sameena Naaz, "Credit Card Fraud Detection using Local Outlier Factor and Isolation Forest " JCSE. International Journal of Computer Sciences and Engineering,Vol.-7, Issue-4, April 2019

[4]Nilsonreport.com. (2019). [online] Available at: https://nilsonreport.com/upload/content_promo/The_Nilson_Report_10- 17-2016.pdf [Accessed 6 May 2019].

[5]S P Maniraj, Aditya Saini, Swarna Deep Sarkar, Shadab Ahmed "Credit Card Fraud Detection using MachineLearning and Data Science" International Journal of Engineering Research & Technology (IJERT), Vol. 8 Issue 09, September-2019

[6 T. Patel and M. O. Kale, ―A Secured Approach to Credit Card Fraud Detection Using Hidden Markov Model, ‖ vol. 3, no. 5, pp. 1576 – 1583, 2014.

[7]V. Vijayakumar, Nallam Sri Divya, P. Sarojini, K. Sonika "Isolation Forest and Local Outlier Factor for Credit Card Fraud Detection System", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-4, April 2020

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