iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VII **Month of publication:** July 2022

DOI: https://doi.org/10.22214/ijraset.2022.45764

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

Credit Card Fake Revealing Analysis and Prediction of Algorithms

Sushma M V[1], Mrs. B R Vatsala[2], Dr. C Vidya Raj[3]
[1]*Information Technology, (Department of Computer science and Engineering)The National Institution of Engineering, Mysore, India*
[2]*Assistant Professor, (Department of Computer science and Engineering), The National Institution of Engineering, Mysore, India*
[3]*Professor and Head, (Department of Computer science and Engineering), The National Institution of Engineering, Mysore, India*

Abstract: ***Credit cards are frequently used in conjunction with the Internet to make payments. The project primarily aims to detect credit card fraud in the real world. Today, our lives have become more and more dependent on online transactions. As technology advances, the number of fraud cases also increases, and ultimately, a fraud detection algorithm needs to be developed to accurately find and eradicate fraudulent activity. Some Machine Learning algorithms can be applied to collect data to address this problem. These machine learning algorithms ultimately optimize the accuracy of the result data. Technique performance is evaluated based on accuracy, f1 score, and precision and recall. Then, processing some of the technique provided attributes identifies the fraud detection and provides the visualization of the graphical model. Classification based algorithms such as logistic regression, random forest and KNN for processing highly unbalanced datasets.***
Index Terms: ***Machine Learning, accuracy, visualization, logistic regression.***

I. INTRODUCTION

As a payment tool, credit cards are widely used by both online and offline by customers due to their convenience and options for borrowing money. However, this convenience has not been without its drawbacks. In recent years, hackers and criminals have become increasingly interested in credit card transactions. In order to use a credit card online, all you have to do is enter the card information, and the physical card does not have to be presented. Some cases may require sending a One Time Password (OTP) as an additional authentication factor. In cases where this is not necessary, such as in the card can be used to make unauthorised purchases in international transactions. Card-Not-Present usage is so named because instead of a physical card, only the card's details are required. There are many ways to steal card information, including card stealing, shoulder surfing, obtaining credit card information, and sniffing web traffic. Cardholders, issuing banks, and merchants are all Fraud victims of credit cards because one of them must accept the burden of scam. In general, it is the cardholder's responsibility to report the bank has the responsibility of detecting fraud. When evidence of fraud is discovered, the bank reverses the credit for the transaction if there is evidence of fraud.

Credit cards are small malleable rectangular-shaped cards give out by a commercial institution, for example a bank, to an authenticated user. In order to purchase goods or services, end users can physically present it at point of sale depots or online e-commerce websites. Credit Card Fraud refers to any illegal use of the card in either one of these customs.

Machine learning-based solutions known as Fraud Detection Systems (FDS) detect fraud in an automated manner used by fraudulent connections can be noticed even before end users report them to credit card corporations. By using such a system, fraudulent transactions can be detected before they are entered into the database, thereby preventing fraud from occurring. In addition to reducing false detections, an ideal FDS would also reduce false detection, which occur when a legitimate transaction is intermittent, Affecting the end user's experience.

A. Problem Statement

To design and develop methods for credit card fraud detection and perform a comparative study of the developed methods.

B. Objectives

Collection of dataset from different networking sites and import the required libraries based on dataset.

1) Data is visualized by plotting graphs and charts by using matplotlib.
2) Split the imported dataset into train data (70%) and test data (30%).
3) Machine learning model is built using algorithms such as Random forest, Logistic regression, Decision Tree.
4) Fitting the trained data to build the Machine learning model.

5) After training the model, models are fitted with test data.
6) Then, predicting and finding out the best accuracy from the model.

C. Problem Description

Online shopping with credit cards resulted in a high number of credit card frauds. Credit card fraud must be identified in this era of digitalization. Monitoring is necessary for fraud detection and evaluating the behaviour of different Estimating, detecting, or avoiding undesirable behaviors by users behaviour. To effectively detect credit card fake, we must first understand the various technologies, algorithms, and types of scam that can be detected with credit cards.

II. LITERATURE REVIEW

The term "fraud" refers to illegal or wrong deception intended to benefit an individual or organization. It is a deliberate act with the intention of obtaining unauthorised financial gain against the law, rule, or policy.

Tittle: Selection of optimal credit card fraud detection models using a coefficient sum apperoch
Author: Suman Arora, Dharminder Kumar
Publication year: 2017

Suman Arora [1] in this paper, Many supervised learning algorithms are applied on datasets that are 70% trained and 30% validation. In this paper we compare logistic regression, and support vector machines (SVM), decision tree, and KNN algorithms notes and techniques i.e. 94.59%, 93.59%, 93.24% and 93.25% respectively. Summary of this paper, Support vector machine(SVM) has the highest true-positive rate with 0.5360, and the decrease in false-negative rate is small, which indicates that the F-score of Decision tree classifier is the lowest.

Tittle: The use of predictive analytics technology to detect credit card fraud
Author: Temitayo Hafiz Kosemani, Shaun Aghili
Publication year: 2016

Temitayo Hafiz Kosemani [2] n this paper, an outline of the fraud detection process is given in the flow chart i.e. There is a deep dive into the acquisition of data, pre-processing, exploratory data analysis, and the approaches or algorithms used. According to the algorithm accuracy, K-Nearest Neighbour (KNN), random tree, AdaBoost, and logistic regression are 96.21%, 94.32%, 57.73%, and 98.24%, correspondingly.

Tittle: Online Credit card fraud detection
Author: You Dai, Jin Yan
Publication year: 2016

You Dai, Jin Yan [3] in this paper, They define an efficient method for conducting fraud detection using a Random Forest classification algorithm. Random forest is classified into two types: tree-based random forest and CART-based random forest. They define in detail, with accuracy of 91.96% and 96.77% respectively. This paper summarizes why the another type is superior than the major.

Title: Credit Card Fraud Detection Using Random Forest
Author: Devi Meenakshi.B, Janani. B, Gayathri. S, Mrs. Indira. N
Publication year: 2019

Devi Meenakshi.B, Janani. B, Gayathri. S, Mrs. Indira. N [5] are authors of Credit Card Fraud Detection Using Random Forest. The assignment is often worried with detecting credit score card fraud withinside the actual world. The phenomenal growth withinside the wide variety of credit score card transactions has these days ended in a big growth in fraudulent activities. The random forest algorithm can improve the accuracy of fraud detection. Classification manner of random wooded area set of rules to research information set and consumer cutting-edge dataset. Finally enhance the accuracy of the end result information. The techniques' overall performance is evaluated the usage of accuracy, sensitivity, specificity, and precision. The processing of a number of the furnished attributes then identifies fraud detection and offers the graphical version visualisation. The techniques' overall performance is evaluated the usage of accuracy, sensitivity, specificity, and precision.

Title: A Machine Learning Methodology for Fraud Detection Using SVM
Author: G. Singh, R. Gupta, A. Rastogi, M. D. S. Chandel, A. Riyaz
Publication year: 2016

G. Singh, R. Gupta, A. Rastogi, M. D. S. Chandel, A. Riyaz [7] re the authors of A Machine Learning Methodology for Fraud Detection Using SVM. Banks have been extra cautious in increasing customer credit, but fraudsters have recently managed to escape with billions of dollars. Because of its convenience and exceptional accuracy, machine learning and data analytics are being used in a variety of fields. In this paper, a comparison of SVM and deep learning in dealing with the credit card fraud detection problem is proposed.

III. PROPOSED SYSTEM

A. System Architecture

The credit card database is first retrieved from the root, and then it is a cleaned and validated on the database, which involves removing reiteration, columns are filled with empty spaces, and altering data. Required factoring or categorizing the variable then data is divided into two parts, a trainings dataset and a testsdataset, after being separated. The primary sample is now anyway divided into test and train datasets. The process of feature extraction involves reducing attributes. Unlike selection of feature, According to the predictive significance of the existing attributes, the existing attributes are ranked, extraction of feature actually changes the attributes. The based on the remaining labelled data, models will be evaluated. The researchers implemented several machine-learning algorithms to classify preprocessed data. They used a random forest model to make their predictions. These algorithms are widely used for text classification tasks. The calculation of models is an main stage in the modelling procedure. It assists in the search for the best model to describe our data. Evaluating model shows using training data is not acceptable in data science since it may easily yield overoptimistic and overfit models.

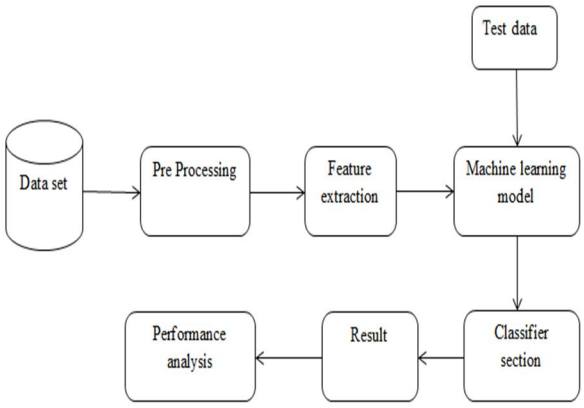


Fig 1: Architecture of the Proposed System

IV. MACHINE LEARNING TECHNIQUES USED

A. Logistic Regression (LR)

Based on one or more features, discovers the best fit variable to predict the probability of the binary response. It is a statistical model An explanation of data and the connection between a single binary dependent variable and a few nominal and ordinal variables. With logistic regression, a set of independent variables can be predicted into binary values(1 / 0, Yes / No, True / False). It is common to use default variables to represent binary or categorical values. It describes relationship between predictors. Predictors are Continuous, Binary, and Categorical. It is applied on the trained model to classify the transactions in the data as fraud or not fraud.

B. Decision Tree (DT)

It is a graph that depicts all possible outcomes of a decision using a branching method. As a supervised learning algorithm, decision trees are used (with a pre-defined objective variable)in classification problems, this method is commonly used. It is applicable to Variables for input and output can be categorical as well as continuous. In this approach, Using the most significant splitter / differentiator in input variables, divide the sample into two or more identical groups.

Types of Decision Tree

1) Categorical Variable Decision Tree: The categorical variable decision0tree consists of a decision tree with a categorical variable as the target variable.
2) Continuous Variable Decision Tree: Whenever the target variable of the decision tree is continuous, it is known as a decision tree with continuous variables.

In the proposed system the dataset has 31 features namely v1…..v28, amount, time, class. There are two steps involved in the process of constructing a decision tree for classification problems: Constructing decision trees based on training data, decide which element belongs to which group using the decision tree for each element. The system is trained and later picks one among the v1…v28 features and starts building the tree the final branching is done in such a way that the class value is checked to find if the transaction is fraudulent or not.

C. K-Nearest Neighbors (KNN)

One of the simplest and most fundamental algorithms for classifying data and A supervised learning approach is used in this model and Application areas include intrusion detection, pattern recognition, and data mining. The fact that it is non-parametric makes it widely applicable to real-life scenarios and does not make any underlying assumptions about the distribution of data. It given some prior data used to classify coordinates into groups according to an attribute. The machine selects one random transaction and classifies it as fraudulent or not fraud, later this classification is compared to the nearest transaction, the process carries on to classify all the available transactions.

D. Support Vector Machine (SVM)

It is a classification and regression challenges can both be addressed using this supervised machine learning algorithm. Most of the time, The method is used to solve classification problems. The SVM algorithm consists of, taking every data item and it can be plotted as a point in n-dimensions and giving its value as a coordinate. Individual observation correspondent are used to evaluate support vectors. As a frontier, SVM does the best job of distinguishing between the two classes (hyperplanes and lines). In SVM the transactions are clubbed into two categories fraud and not fraud , some transactions exits which are difficult to classify hence the class value of such transactions is used to classify them.

The SVM algorithm is best understood by centering on its fundamental type, the SVM classifier. The SVM classifier is designed to create a hyper-lane in an N-dimensional space that splits data points into multiple groups. This hyperplane, however, is selected based on margin, as the hyperplane with the greatest margin between the two clasification is evaluated.

V. METHODOLOGY

A. Library Modules

1) Numpy

NumPy is a Python package. It is an acronyms for 'Numerical Python.' It is a library that includes multidimensional array objects as well as array processing techniques. NumPy package was created by integrating Numarray functionalities into the Numeric package. Several people have contributed to this open source project. NumPy may be used by a developer to perform the following tasks:

- operations on arrays that are logical and mathematical.
- Shape-manipulation methods and Fourier transformations.
- Linear algebra-related operations. Linear algebra and random number generation have built-in functions in NumPy.

2) PANDAS

Using Pandas as a tool for handling data structures and data analysis is easy because Pandas is an open-source, BSD-licensed Python library. There are many university and economic disciplines that use Python with Pandas, including finance, economics, statistics, analytics, etc. In this lesson, Python Pandas can be used for a wide variety of purposes, and we will learn how to use them effectively during this course. There are three types of data structures that Pandas works with: Series, DataFrames, and Panels.

- Clean the data by eliminating missing values and filtering
- rows or columns based on specified criteria.
- Use Matplotlib to visualise the data. Create graphs using bars, lines, histograms, bubbles, and other elements.
- Restore the cleaned and modified data to a CSV, other file, or database.

3) MATPLOTLIB

This is a collection of command-style techniques that make Matplotlib behave like MATLAB. Each pyplot function modifies the diagram in some way. For example, create characters, plot areas within characters, draw specific lines within plot areas, style plots with labels, and etc. Matplotlib maintains a variety of states between each function call.pyplot, tracking the current figure, plot area, etc., and the plot function is oriented to the current axis.

4) SKLEARN

SVMs, gradient boosting, k-means, random forests and other machine learning algorithms are available in Scikit-learn, and DBSCAN are among the regression, classification, and clustering algorithms included in it. In addition to Python Numpy and scripts, it is designed for use with Python. As a part of the Google Summer of Code program, the scikit-learn project began (also known as GSoC).

B. Credit Card Dataset

Kaggle is a data analysis platform that provides datasets for data analysis. There are 31 columns total in this dataset, but sensitive information is protected by assigning 28 of them that the names v1 to v28. Also included are columns for Time, Amount, and Class. The Time field displays the interval between the first and second transaction. The amount of money exchanged is called the amount. Class 0 denotes a genuine transaction, whereas Class 1 denotes a fraudulent one.

1) Understanding the Data

The primary data gathered from web sources is presented in the form do statements take, numerals, and subjective phrases. There are errors, omissions, and discrepancies in the metadata. After carefully reviewing the completed surveys, adjustments are required. The following steps are involved in primary data processing. It is necessary to categorize a massive amount of raw data obtained from a field survey for comparable information across respondents. The data Includes credit score card transactions made through cardholders in September 2013. This dataset carries 492 instances of fraud from 284,807 transactions that befell in days. The dataset is drastically imbalanced, with high quality transactions (illegal) accounting for 0.172% of all transactions. Basically, The input variables are numerical variables that result from PCA. Unfortunately, project can't provide the original data or more information about it because of confidentiality concerns. The only features not changed by PCA are 'time' and 'amount'. The features V1, V2, and V28 were derived using PCA. There is a feature called "time" that indicates the number of seconds between the initial transaction and the subsequent transactions in the dataset. The amount displayed in the "Amount" feature can be used for example-dependent, cost-sensitive learning. The "Class" feature has a value of 1 when fraud is detected and 0 when it is not detected.

2) Data pre-processing

Machine learning requires data and models in order to work. When collecting data, make sure to include enough features (aspects of the data that might help with prediction, such as the surface of a building used to predict its price) Your learning model needs to be properly trained. Preprocessing raw data is The preparation of a dataset for machine learning. The first step in building a machine learning model is to establish a baseline. Data from real-world sources often contain noise, missing values, and may be in a format that cannot be directly applied to machine learning models. The process of converting raw data into an useable format is known as data preprocessing. Real-world data is frequently inadequate, inconsistent, and/or missing in specific behaviours or trends, and it is filled with errors. A tried-and-true method of solving these issues is data preprocessing.

Data preprocessing is a necessary function for cleansing the information and getting ready it for a system mastering model, System mastering models are improved in accuracy and performance as a result.

3) Data Visualization

A graph is a visual representation of data and information. Data visualization tools are helpful for viewing and analyzing data trends, outliers, and patterns.

It is used because:

- Data visualization identifies trends in data
- Data visualization gives the data a perspective
- Visualization of Data places data in the appropriate context.
- Time Savings using Data Visualization
- The Data Story is Told Through Data Visualization

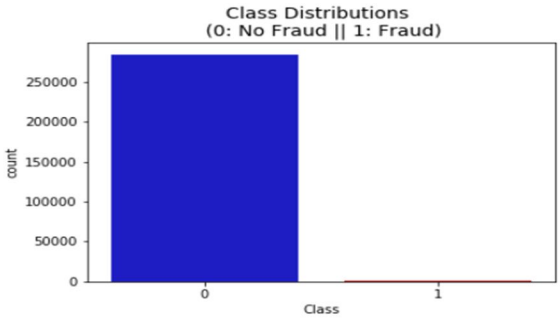


Fig 2: Visualization of Data Distribution.

a) Scaling and Distributing

The purpose of feature scaling is to standardize the independent variables used in data within a specific range. The intention of using feature scaling is to address drastic variations in magnitude, value or units. If feature scaling is not performed, machine learning algorithms can result

In mathematics, a distribution is just a collection of values or scores on one or more variables. These rankings are typically organised from lowest to biggest and then graphically shown.

In distribution, the number of different possible ways the data can be presented, the percentage of different data points, and the identification of outliers are the main terms. So, data distribution is the process of organising and displaying valuable information utilising graphical ways.

The Time and Amount columns should be scaled first. Time and amount should be scaled in accordance with the additional columns. As part of our efforts to support our computers in understanding how patterns define whether a transaction is fraudulent or not, a subset of the data frame was created with an equal number of fraudulent and non-fraudulent transactions. In order to test the results of the proposed filtering algorithm, a sub-sample consisting of a data set with a 50/50 ratio between fraudulent and non-fraudulent transactions was selected. The proportion of fraudulent and Non-fraudulent transactions in the sub-sample will be equal.

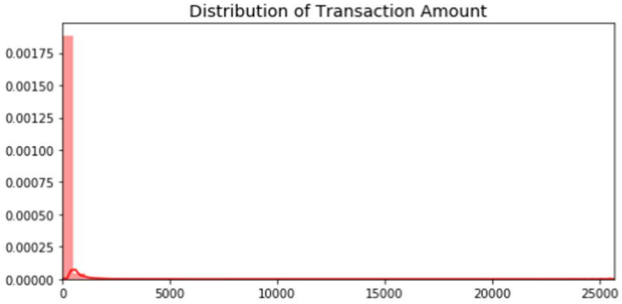


Fig 3: Distribution of Transaction Amount

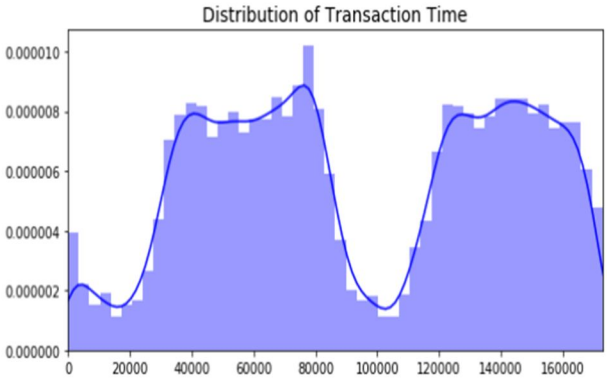


Fig 4: Distribution of Transaction Time

As a result of the heavy imbalance in the original data frame, the original data frame will result in the following problems:

- Over fitting: The classification models in this project assume there are no-longer any frauds! Fraud needs to be detected by our model with confidence.
- Inaccurate Correlations: Although we cannot yet definitively identify what the "V" shapes refer to, it may be useful to consider their binary relationship with the classifier's final result (Fraud or No Fraud)from the standpoint of unknown data. If a large percentage of "V" shapes are observed.

b) Splitting the Data

When data is split, A subset of it is broken down into two or more parts. It is typical to split the data into two parts to evaluate or test the data and train the model in the other part. Training sets are commonly used to determine various extensity or to compare model performance. After training, the testing dataset is accustomed. To ensure that the final model works properly, the training and test data are compared. Data can be divided in no universal manner or based on any universal metric; It may depend on the size of the original data repository or how many predictors are included in the predictive model. Organizations and data designers Depending on the methods used for sampling, data may be split into different categories.

Rather than using either of the techniques to create testing sets, models should be tested using the original testing set. As the main objective is to detect patterns in data that were under sampled and oversamples, and then test the model with the original data, the model is fitted either with the results of under sample and over sample or with the results of over sample and under sample.

Generally, the original data in a machine learning prototypical is divided into three or four sets. The training set, development set, and testing set are the three most commonly used sets:

- During the training process, a model is trained using the training set. Any parameters in the model should be optimised based on what is observed and learned from the training set.
- The Developing sets of examples are a way of changing the learning process parameters. A model authentication set can also be referred to as a cross-authentication set. Using this data set, you can rank the model's accuracy and choose the most appropriate model based on its accuracy.
- During testing, the testing set is compared with the previous sets of data to determine if the final model accurately predicts the results. The testing set serves as an analysis of the chosen algorithm and mode.

C. Random Under Sampling and Over Sampling:

Random Under Sampling is an approach to removing data from our dataset in order to develop a dataset that is more balanced and therefore prevent overfitting.

The steps to take can vary but may include:

1) To determine the nature of class imbalance, run the "value_counts()" function on the class column, which will show the count of the different labels.

2) To determine how many fraud transactions took place in the first day, these instances must be subtracted from the Count of totals of non-fraud transactions to reach the total number of fraud transactions (assuming a 50/50 ratio of fraud and non- fraud)

3) Using this technique, we were able to obtain a 50/50 class ratio subsample of our data frame. As a second step, Every time this script runs, we will shuffling the data to determine if models are accurate.

a) Distributing and Correlating

- Equally Distributing Classes

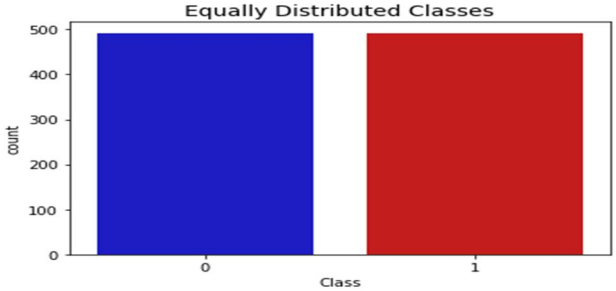


Fig 5: Equally Distributing Classes

- *Correlation Matrices*

Correlation matrices present the covariance coefficients for multiple variables in a table of data. An example of a matrix is an illustration of how all possible pairings of values in a table are related. In addition to identifying and visualizing patterns in large datasets, it is a powerful tool for summarizing large datasets.

In order to understand the data, correlation matrices are crucial. There are some factors that heavily inspiration whether a particular transaction is fraudulent. In order to determine whether negative or positive transaction characteristics are associated with fraud-related features, it is critical to use the correct data around in demand for that analysis.

Each variable is represented by a row and a column in a correlation matrix. Throughout the table, the correlation coefficient can be found in each cell.

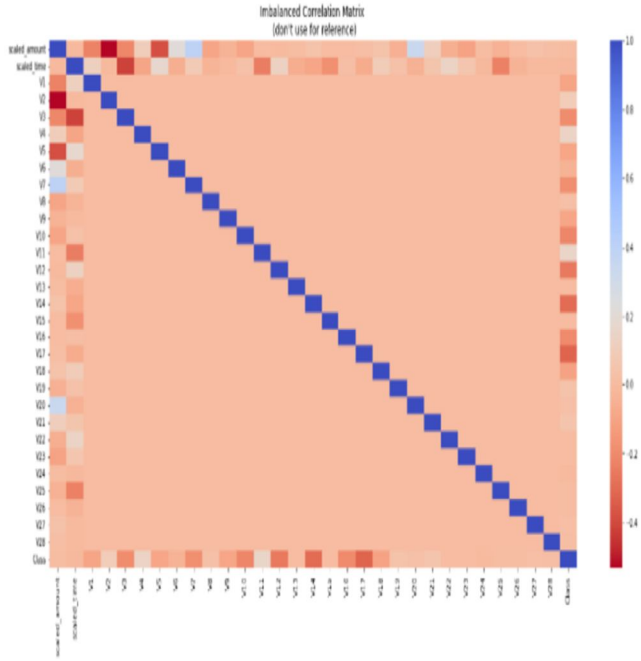


Fig 6: Correlation Matrix

Better accuracy may be gained utilizing the Outlier Data mining approach, as previously stated. Because the bias and absolute values have been discovered and eliminated, better results may be produced; so now, our objective now is to find and remove the extreme outliers. First, a reference model will be created to visualize the relationship between the variables.

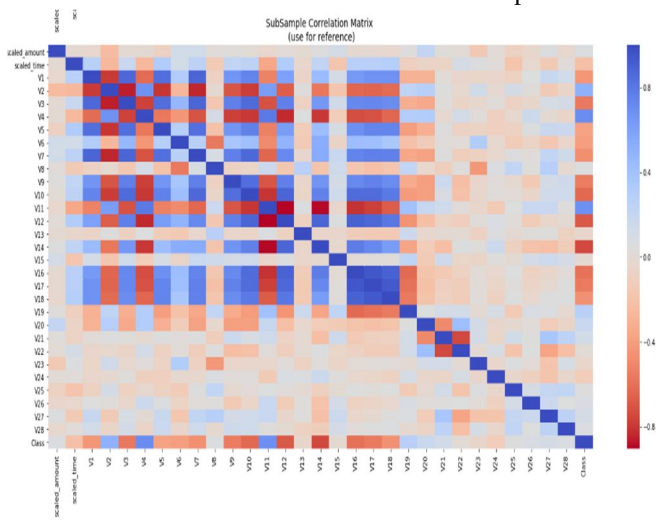


Fig 7: Sample Correlation Matrix

b) Anomaly Detection

Anomaly Detection is By detecting anomalies, features with a high correlation with classes can be removed from feature sets. The accuracy of the models will be enhanced the positive impact the accuracy as a result.

Explanation

Visualize Distributions: visualization such as this one will be used to eliminate some of the outliers from the distribution of features. A Gaussian distribution is present only in V14, while a Non-Gaussian distribution is present in V12 and V10.

Defining the threshold: We will multiply the number with the IQR (the lower number will have fewer outliers removed), and then determine the upper and lower thresholds by substrating q25 - threshold (lower extreme threshold) and adding q75 + threshold (upper extreme threshold).

Provisional Dropping: A provisional dropping in which the instances are removed As long as both extremes of the threshold are exceeded.

Boxplot Representation: There has been a dramatic reduction in the number of "extreme outliers" as shown in the boxplot.

c) Dimensionality Reduction and Clustering (t-SNE)

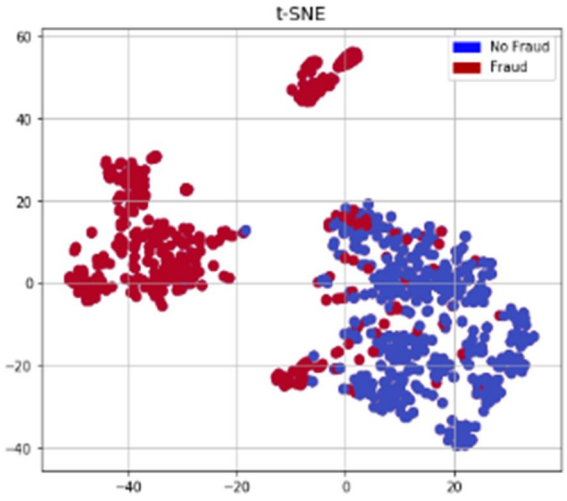


Fig 8: Clustering

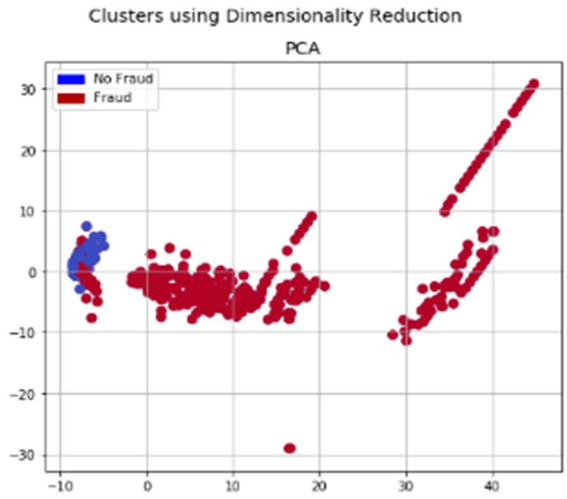


Fig 4.8 Clustering of PCA

Understanding t-SNE

- Clustering of the fraud cases and non-fraud cases can be done fairly accurately using an algorithm.
- A small subsample doesn't matter, t-SNE can detect clusters pretty accurately in all scenarios, regardless of the subsample size
- Taking this into account, we can expect future predictive models to do a fairly good job of separating fraud-related cases from non-fraud-related ones.

d) Classifiers(Over Sampling)

Assess the effectiveness of various fraud detection algorithms and decide which classifiers should be used. Clearly separating the features from the labels and dividing the training and testing data into training and testing sets is essential before proceeding.

- *Explanation*

➢ Classifiers based on logistic regression are typically more accurate than those based on other three types of classification.
➢ The GridSearchCV algorithm is used to determine what parameters offer the best predictive score for the classifiers.
➢ Its Receiving Operating Characteristics (ROC) score is the highest, meaning that Logistic Regression separates fraud transactions from non-fraud transactions quite accurately.

- *Learning Curves*

➢ A model that is overfit is more likely to be overfit if the difference between training and cross validation scores is small (high variance).
➢ If we get low scores in both training and cross-validation, it indicates that our model is underfit (high bias)
➢ Cross-validation and training sets show a better performance for the Logistic Regression Classifier.

D. Oversampling with SMOTE

SMOTE is an abbreviation for Synthetic Minority Over-sampling Technique. SMOTE, unlike Random Under Sampling, The system generates new synthetic points to ensure that all classes are equally balanced. There are many ways to resolve problems caused by “class imbalance”, but this is one of them.

VI. RESULT AND PERFORMANCE ANALYSIS

A. Confusion Matrix

1) Positive/Negative: Type of samples (label) ["No", "Yes"] True/False: The model rightly or erroneously categorised.
2) True Negatives (Top-Left Square): The following counts the number of results that were marked “No”. (No Fraud Detected) sample.
3) False Negatives (Top-Right Square): The following counts the number of results that were marked "No"(No Fraud Detected) sample.
4) False Positives (Bottom-Left Square): The following counts the number of results that were marked "Yes" (Fraud Detected) sample.
5) True Positives (Bottom-Right Square): The following counts the number of results that were marked "Yes" (Fraud Detected) sample.

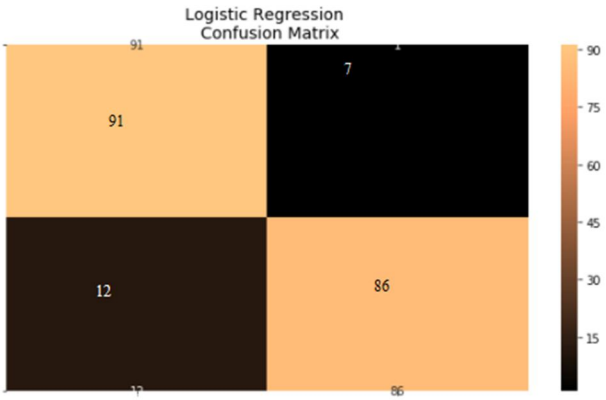


Fig 10: Confusion Matrix Logistic Regression

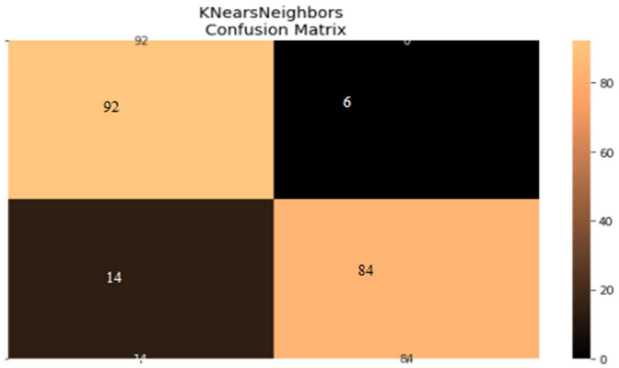


Fig 11: Confusion Matrix KNN

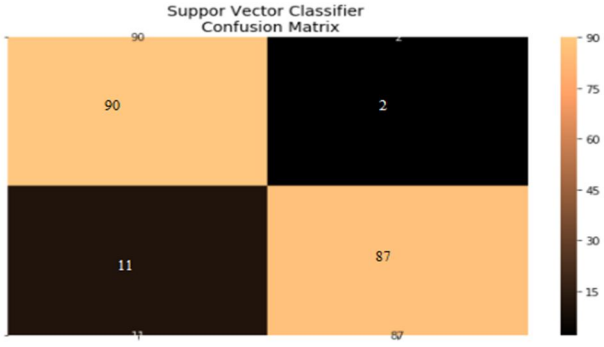


Fig 12: Confusion Matrix SVM

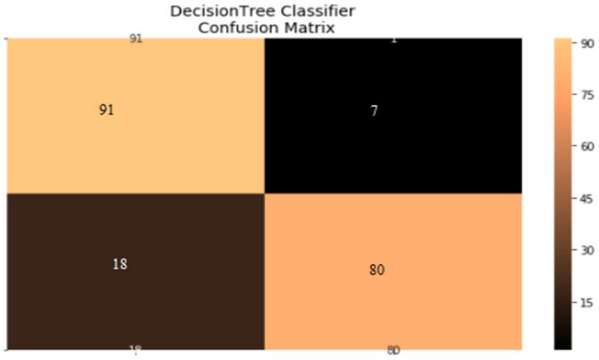


Fig 13: Confusion Matrix DecisionTree

It prints out how many false positives were detected and compares them with the actual ones. In this way, accuracy scores and precision scores of algorithms are calculated.

TABLE I. TABLE OF ACCURACY OF ALGORITHM

Sl.No	Algorithm	Accuracy
1.	Logistic Regression	0.94
2.	K Nearest Neighbours	0.93
3.	Support Vector Classifier	0.93
4.	Decision Tree Classifier	0.93

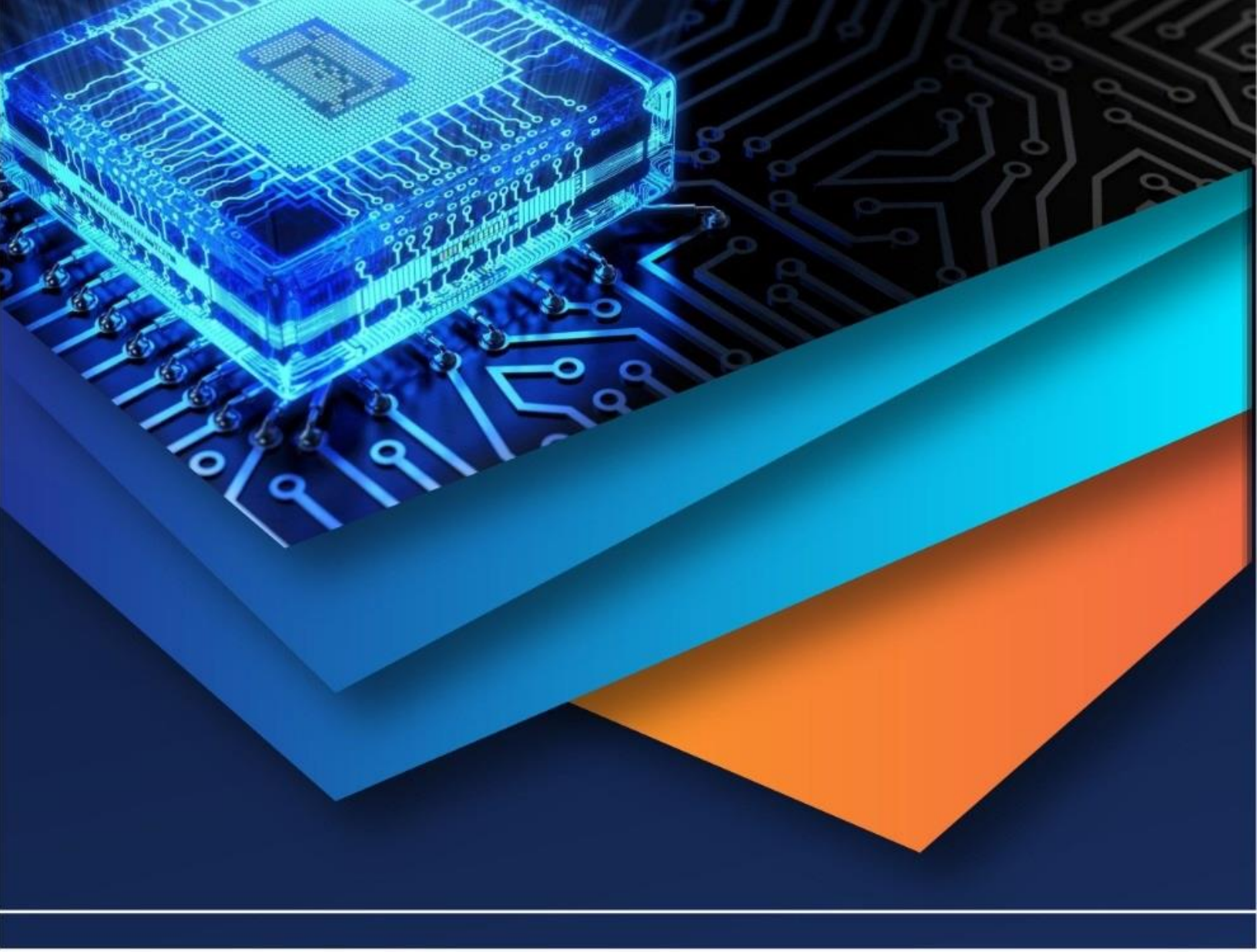
Table 1. Performance of the algorithm

VII. CONCLUSION

It is particularly difficult to detect credit card fraud because of the imbalance between the normal and fraudulent transactions. The purpose of this is to detect credit card fraud, Logistic regression, Decision Tree, KNN, and SVC techniques were used. Precision score, F1 score, Accuracy, and recall are used to evaluate the proposed system's performance. Based on logistic regression, decision tree, KNN, and SVC, the accuracy is 94.05%, 91.41%, 92.73%, 93.79%, respectively. The high percentage of accuracy is expected since valid transactions are much more numerous than genuine ones. The logistic regression classifier and support vector classifier are better than the KNN and decision tree based on a comparison of all four methods.

REFERENCES

[1] Credit card Fraud Detection based on Machine Learning Algorithms- Heta Naik and Prashasti Kanikar - International Journal of Computer Applications - March 2019
[2] Credit Card Fraud Detection Using Random Forest - Devi Meenakshi, Janani, Gayathri and Mrs. Indira - International Research Journal of Engineering and Technology (IRJET) - March 2019
[3] Credit Card Fraud Detection using Machine Learning and Data Science -Aditya Saini, Swarna Deep Sarkar Shadab Ahmed and S P Maniraj - International Research Journal of Engineering and Technology (IRJET) - September-2019
[4] D. Tanouz, R Raja Subramanian, D. Eswar "Credit Card Fraud Detection Using Machine Learning" International Conference on Intelligent Computing and Control Systems (ICICCS 2021)
[5] A Review of Credit Card Fraud Detection Using Machine Learning Techniques- Nadia Boutaher Amina Elomri, Noreddine Abghour, Khalid Moussaid and Mohamed Rida, International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications- 2020
[6] (D. Tanouz, 2021)Credit Card Fraud Detection System Using Machine Learning- J. Devi Sushma, M. Hemanthkumar, P.H.V.Raviteja-International Journal of Research in Advanced computer Science Engineering (ijracse)-2019
[7] Supervised Machine Learning Algorithms for CreditCard Fraud Detection: A Comparison- Samidha Khatri, Aishwarya Arora, Arun Prakash Agrawal- International Conference on Cloud Computing, Data Science & Engineering-2020



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY