International Journal on Applications in Information and Communication Engineering
Volume 8 : Issue 2 : June 2022, pp 17 – 22 www.aetsjournal.com ISSN (Online) : 2394 - 6237

CREDIT CARD FRAUD DETECTION USING DATA SCIENCE AND MACHINE LEARNING TECHNIQUES

S LEKHANA , A S DADAPEER , G JAGAN , V HAREESH KUMAR REDDY

Abstract — In this paper, we propose defining a measure of efficient technique for detecting fraud in credit card usage. A credit card is issued by a bank or financial services company and allows cardholders to borrow funds to pay for goods and services at merchants who accept cards. Because everything is digitalized, there is a risk of card misuse and account holders losing money; As a result, credit card companies must be able to detect fraudulent credit card transactions. Data science and machine learning approaches can be used to tackle this sort of challenge. It is concerned with the modelling of the dataset using machine learning with Credit Card Fraud Detection. Modeling historical credit card transactions with data from those that turned out to be fraudulent is an important component of machine learning. After then, the developed model is utilised to identify whether or not a new transaction is fraudulent. The purpose is to assess whether or not the fraud took place. Before applying a machine learning algorithm to a dataset, the first stage comprises assessing and pre-processing the data. credit card dataset to determine the parameters of the algorithm and calculate their performance metrics.

Keywords— Credit card, Fraud detection, Outlier, Random Forest, Naïve algorithm

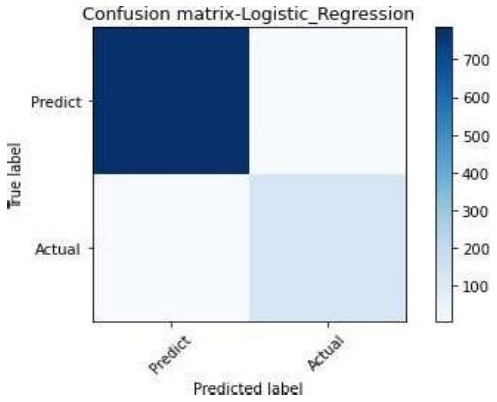
I. INTRODUCTION

The training's goal is to build a classification model that can detect whether a transaction is fake. A dataset of previous credit card cases is compiled and used to train the machine to recognize the problem. The first step is data analysis, which involves analyzing each column and taking the necessary measurements for missing values and other types of data. Outliers and other values with little influence are dealt with. Machine learning algorithms are used on training data to learn patterns from the data.

Fraud detection may be done in any way, with the dataset determining when to employ which method. Several supervised algorithms have been employed to identify credit card fraud inrecent years. Algorithms in Supervised Learning learn from labelled data. The algorithm selects which label should be assigned to new data based on pattern and linking the After understanding the data, the algorithm determines which label should be given.

A. Maintaining the Integrity Through Logistic Regression

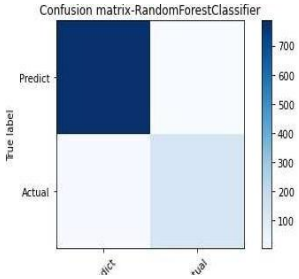
It's a statistical technique for assessing a data collection in which one or more independent factors influence the outcome. The result is measured using a dichotomous variable (in which there are only two possible outcomes). The purpose of logistic regression is to determine the best model to represent the connection between a group of independent (predictor or explanatory) variables and a binary feature of interest (dependent variable = response or outcome variable). In logistic regression, the dependent variable is a binary variable that contains data coded as 1 or 0. (yes, success, etc.)



S Lekhana , Department of CSE, Madanapalle Institute of technology & Sciences, Madanapalle,AP., INDIA.
(Email: 18699A0552@mits.ac.in)
A S Dadapeer , Department of CSE, Madanapalle Institute of technology & Sciences, Madanapalle,AP., INDIA
(Email: 18699A0507@mits.ac.in)
G Jagan , Department of CSE, Madanapalle Institute of technology & Sciences, Madanapalle,AP., INDIA
(Email: 18699A0516@mits.ac.in)
V Hareesh Kumar Reddy , Department of CSE, Madanapalle Institute of technology & Sciences, Madanapalle,AP., INDIA (Email: 19690A0503@mits.ac.in , lekhana.sandra@gmail.com)

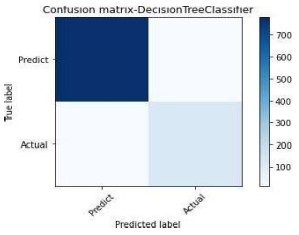
B. Random Forest classification for efficiency

Random forests generate the class that represents the mode of the classes (classification) or the mean prediction (regression) of individual trees. Decision trees have a propensity to overfit their training set, which is countered by random choice forests. A random forest is a collection of trees .



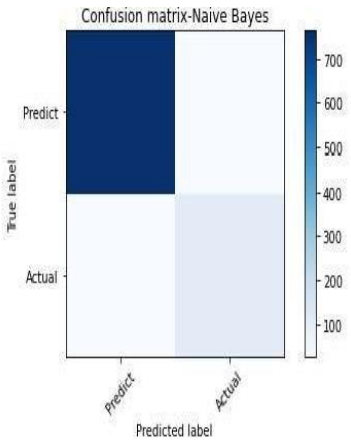
C. Decision tree-based fraud detection

It breaks down a large data set into smaller and smaller chunks while simultaneously building a decision tree.A leaf node indicates a categorization or decision, while a decision node has two or more branches.



D. Naïve Bayes Approach for detection

The Naive Bayes algorithm is a straightforward method for predicting the likelihood of each feature belonging to each class. If you wanted to model a predictive modelling issue probabilistically, this is the supervised learning strategy you'd use.



II. LITERATURE SURVEY

Fraud trends developed over time, creating new types of fraud, making it a hot topic among scholars. The rest of this section goes through individual machine learning algorithms, machine learning models, and fraud detection systems that have been employed in fraud detection. The issues that arose throughout the study have been investigated in order to develop an effective machine learning model in the future. Past researchers discovered several issues with fraud detection after analysing various detection algorithms. It employs a Back Dissemination network to correct values that have been proven to be incorrect. All of these procedures have major challenges, such as diminishing accuracy levels and deficient capabilities.

In [1] According to the author, rising trends in financial transactions using credit cards also encourage fraud, resulting in billions of dollars in global losses.To assess fraud detection activities that need analysis of a large amount of transaction data is accessible. The proposed study provides a condensed overview of several strategies for classifying fraud transactions from diverse datasets and alerting users to such transactions. Online transactions have been the most frequent medium for financial transactions in recent decades.

In [2] It has been mentioned that credit card fraud has become one of the most prevalent issues in recent years. Individuals who use credit cards, as well as retailers and banks, have all suffered significant financial losses. This study compares multiple machine learning-based fraud detection approaches utilizing performance measures such as accuracy, precision, and specificity. In addition, the research presents an FDS based on the supervised Random Forest method. The accuracy of identifying credit card fraud is improved with this suggested solution. Furthermore, the suggested system ranks the alert using a learning to rank technique, successfully addressing the problem of idea drift in fraud detection

In [3] The study primarily focuses on detecting credit card theft in the real world. For the qualifying data set, we must first gather credit card data sets.

Then, to test the data set, run queries on the user's credit card. After using the previously assessed data set and supplying the new data set, the random forest algorithm classification technique is used. Then the processing of a few characteristics will be applied, so that impacting fraud detection may be detected in reading the representation of the graphical model. The efficiency of the procedure is determined by its accuracy, adaptability, and specificity, as well as precision.

In [4] Author proposed a wat to assess fraud detection activities that need analysis of behavior/abnormalities in the transaction dataset to discover and disregard the suspected person's unwanted activity, a large amount of transaction data is accessible. The proposed study provides a condensed overview of several strategies for classifying fraud transactions from diverse datasets and alerting users to such transactions. Online transactions have been the most frequent medium for financial transactions in recent decades.

In [5] Author made a clear discussion on Companies seek to provide their consumers with more and more services Customers may now purchase essential items online. Criminals can steal any cardholder's details and use it for online transactions until the cardholder calls the bank to get the card blocked. This article demonstrates the various machine learning strategies that are used to identify this type of transaction. According to the findings, CCF is a big concern in the financial industry that is becoming more prevalent over time. More and more businesses are migrating to an online method that allows clients to conduct transactions online. Criminals can use this chance to steal other people's information or credit cards in order to conduct online transactions..

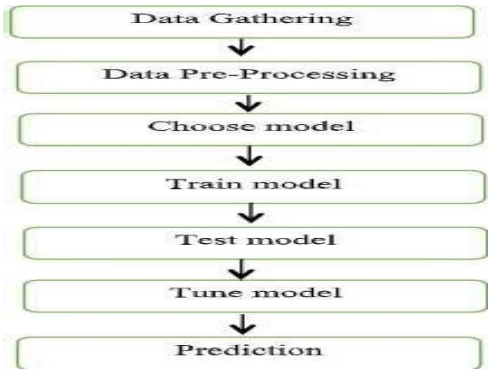
In [6] Author has that Users may buy consumable durable things online and move money from one account to another by utilizing their credit cards. By using phishing, Trojan viruses, and other methods, the fraudster is able to determine the specifics of the user's transaction activity and engage in illicit actions with the card.This will assist to increase card transaction security in the future.. Many credit card customers are losing money and sensitive information as a result of these fraudulent operations. We explored several fraudulent detection and control strategies in credit cards in this research, and it will be useful in the future to strengthen fraudster security and avoid criminal actions.

In [7] Since the algorithms mentioned in the study are concerned with machine learning approaches, the scope of the referred paper may be considered. It has focused on ways for gathering data from social media and structuring them in terms of big data models, as well as addressing the field's problems. They conducted a thorough investigation on Apache Spark, which employed fuzzy clustering logic for massive data processing.

To detect and avoid credit card frauds, a variety of competent techniques such as arrangement orientation, device learning, neural networks, artificial intelligence, and fuzzy logic are used. In today's world, fraud is one of the leading causes of large-scale corporate losses, affecting not just merchants but also individual clients. As a result, there exist several ways for detecting such scams. Initially, a clustering approach was used to identify legal and illegal operations using data cauterization of factor value regions. In addition, a Gaussian mixture model is used to model the potential thickness of a credit card operator's prior performance, so that the probability of current actions may be meant to detect any abnormalities from historical behavior.

Exploration data analysis of variable identification:

- Loading the given dataset
- Import required libraries packages
- Analyze the general properties
- Find duplicate and missing values
- Checking unique and count value



Method of Outlier detection with feature engineering:

- Pre-processing
- Splitting and training Comparing the Decision trees

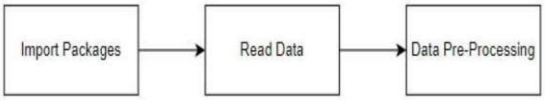
Comparing algorithm to predict the result:

- Based on the best accuracy

The main goal is to use a machine learning method to detect Fraud Prediction, which is a typical text classification problem.

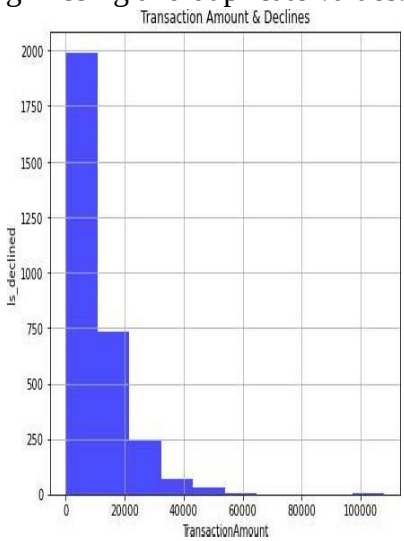
A. Data Pre-Processing

While tweaking model hyper parameters, a sample of data is employed to offer an impartial evaluation of a model fit on the training dataset. When competence on the validation dataset is included into the model architecture, the evaluation becomes increasingly skewed. Although it is only used on a regular basis, the validation set is used to test a model .



B. Data validation/Preparing/Cleaning:

Importing library packages and loading the specified dataset. To investigate variable identification based on data form and type, as well as analyzing missing and duplicate values.

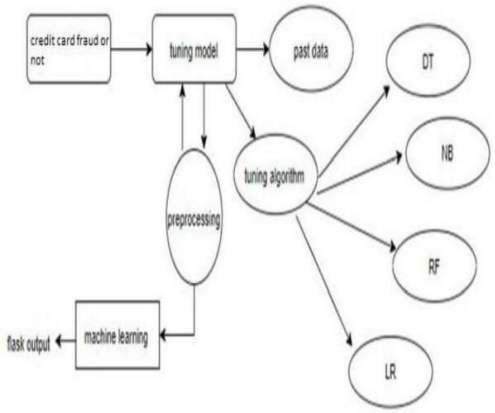


After training your model that is used to measure model skill when tweaking models and processes for making the greatest use of validation and test datasets while evaluating your models. Cleaning and preparing data by renaming the dataset and removing columns, among other things.

This might be useful for spotting trends, faulty data, outliers, and other things while exploring and getting to know a dataset. Data visualizations may be utilized to convey and show crucial links in plots and charts that are more visceral and intuitive than measurements of association or significance with a little topic expertise.

C. Comparing Algorithm with prediction in the form of best accuracy result:

The performance characteristics of each model will vary. You may gain an idea of how reliable each model is on unseen data using resampling approaches like cross validation. It must be able to use these estimates to select one or two of the best models from the set you've created. The same idea applies to model selection .



Prediction result by accuracy: True Rate= TP / (TP + FN)

False Rate = FP / (FP + TN)

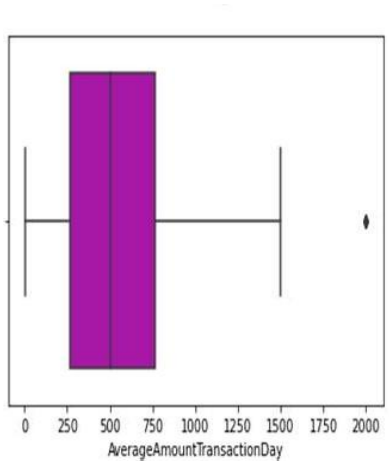
Calculation:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision: The proportion of positive predictions that are correct.

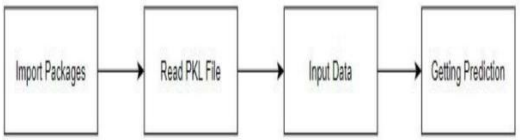
Precision = TP / (TP + FP)

***Recall*:** The percentage of observed positive values that were accurately anticipated. (The percentage of real defaulters predicted properly by the model)



D.FlaskThe Web Framework

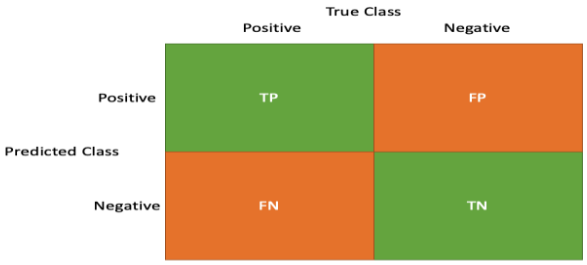
It was created by Armin Ronacher for pocoo. Extension framework features include object-relational mappers, form validation, upload handling, a slew of open authentication techniques, and a slew of other common technologies. Armin Ronacher of Pocoo, a worldwide association of Python aficionados founded in 2004, invented Flask. Because it doesn't require any tools or libraries, it's categorised as a microframework .



III. RESULTS AND ANALYSIS

Cardholder transaction data over the previous 10 years was picked, comprising 16,584 transactions, 15,135 of which were non-fraudulent and 1,449 of which were fraudulent. Cardholders and fraudsters have quite different spending habits. Account and transaction data from cardholders might disclose a lot about their spending habits As a result, object properties for account and transaction records are chosen. It's difficult to assess fraudulent transactions just based on transaction-related information. Every sample contains 51 different character .

The fraud detection system for real-time detection of credit card fraud is made up of three primary components: API MODULE, FRAUD DETECTION MODELS, and WAREHOUSE. At the same time, all of the components are working to detect fraud. The four types of fraudulent transactions are: (Frauds arise as a result of Risky Transactions).



Classification report of Random Forest Results:

	precision	recall	f1-score	support
0	0.99	1.00	0.99	789
1	0.98	0.95	0.96	134
accuracy			0.99	923
macro avg	0.98	0.97	0.98	923
weighted avg	0.99	0.99	0.99	923

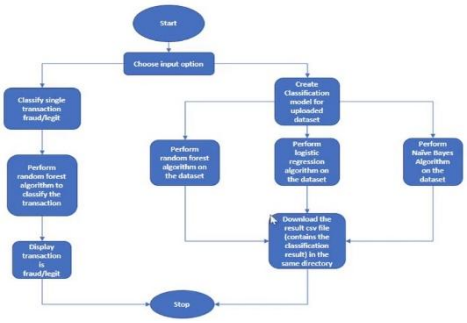
Confusion Matrix result of Random Forest Classifier is:
[[786 3]
[7 127]]

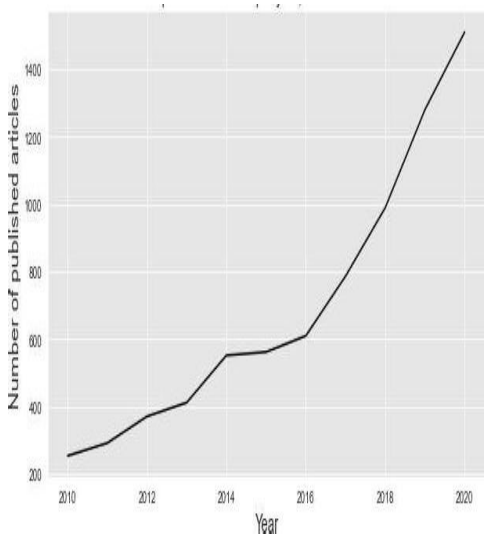
Sensitivity : 0.9961977186311787

Specificity : 0.9477611940298507

Cross validation test results of accuracy:
[0.97723577 0.99349593 0.9804878 0.98536585 0.98211382]

Accuracy result of Random Forest Classifier is: 98.3739837398374





IV. CONCLUSIONS

Within this broadsheet, an analysis of a credit card theft using AI equations has been added. Similar template mos using DL, NB and SVM have been used in experimental evaluation. Free available Master card knowledge index been used evaluation using singular (normal) model-half breed model using Boost and lion's share casting a ballot exhibition ration because it deliberates the valid and vital bogus optimistic bad results predicted. Beside all these an metric evaluation methods for the performance evaluations the proposed algorithms the algorithm can predict the fraud credit card business up to some level whereas the possibility fraud occurrences in credit card business are in through many intermediate channels. Construction of INSI categorized data between spurious data and finding dependencies among them in all aspects is difficult. The conclusion states that the POST proposed technique is limited to some extent only.

V. FUTURE WORK

For future works, the strategies concentrated in this broadsheet will be stretched out to web based book learning copies. In expansion, other web based ignorance model will be examined. The application of web based wisdom will enable quick location of falsification case, conceivably continuously. This will help distinguish and forestall untrustworthy interactions before they happen, which lessens the measures of disasters assimilated each day in the budgetary part.This application can help to find the Prediction of credit card fraud or not

- The cloud model will be used to anticipate credit card fraud.
- In an Artificial Intelligence environment, to make the job easier to implement.

REFERENCES

[1] Wang Xi. Prediction China Trial. Apr. 2008, pp. 74.Computersettained conference 2019.
[2] J.Laurikkala, M Juhola, E Kentala. Informal Identification of Outilers in Medical Data[C]. In:5th International Workshop on Intelligent Data Analysis in Medicine and Pharmacology,(IDAMAP-2000),2000.
[3] **R.SenthamilSelvan** '' Design and Analysis of Spawn Protocol for Secure Communication in VANET'' on IJET (UAE) Volume 7, Issue 4.19, 2018, 360-365, E-ISSN: 2227-524X
[4] Han J W,KamberM.Data Mining: Concepts and Techniques. Beijing: Higher Education Pr. and Morgan Kaufmann Publishers, 2007
[5] Lu Shenglian, Lin Shimin. Research on Distance- based Outliers Detection. Computer Engineer and Applications.Vol.33, 2004, pp.73- 76 Lu Shenglian, Li