EFFECTIVE DETECTION OF CREDIT CARD FRAUD USING LOGISTIC REGRESSION, DECISION TREE AND MACHINE LEARNING TECHNIQUES

[1]R Uttam Sai, [2]A.Mahesh, [3]N.Ashwini, [4]Sunketi Sravani Reddy
[1,2,3]Assistant Professor, [4]UG Student, [1,2,3,4]Department of Computer Science Engineering, Visvesvaraya College of Engineering & Technology, Hyderabad, India.

Abstract
Credit card theft has increased as e-commerce has expanded. Banks are having a harder time identifying credit card fraud as a result of the industry's explosive growth. One of the most crucial tasks that machine learning performs in preventing credit card fraud is the identification of fraudulent purchases. Banks estimate these types of transactions using a range of machine learning techniques, building on historical data and adding special characteristics to boost prediction accuracy. The strategy for sampling the data set, the choice of variables, and the detection process are all crucial to the effectiveness of credit card fraud detection. The usefulness of decision trees, random forests, and logistic regression in identifying credit card fraud is examined in this study. A dataset of 2, 84,808 credit card transactions from a European bank collected using kaggle. It defines fraudulent transactions as "positive class" and valid purchases as "negative class," but the data set is considerably skewed, with just about 0.172% being fraudulent and the rest being legitimate because to theauthor's frame interpolation efforts, the dataset now contains 60% fraudulent and 40% valid transactions. The dataset is subjected to each of the three approaches, and the output code is written in the programming R. Several metrics, including as sensitivity, specificity, accuracy, and error rate, are utilized to analyze the efficacy of the approaches in relation to the aforementioned criteria. The accuracy of the logistic regression, decision tree, and random forest classifiers was 90.0, 94.3, and 95.5%, respectively. When evaluated to logistic regression and decision trees, the Random forest outperforms both.

Introduction
In order to make a transaction without paying or to withdraw money from an account without authorization, credit card fraud encompasses theft and fraud that is committed at the moment of payment using a credit card. Credit card fraud is frequently linked to identity theft. According to data released by the Federal Trade Commission in the United States, identity theft climbed by 21% in 2008 after being mostly steady throughout the middle of the millennium.
Both the total number of ID theft incidents and the percentage of complaints involving credit card fraud decreased. In 2000, there were nearly 13 billion transactions per year, of which 10 million (or one in every 1300) were fraudulent. Additionally, 5% of all active accounts each month were bogus (5 out of every 10,000). Even though only one-twelfth of all transactions are currently monitored by fraud detection systems, the resulting losses amount to billions of dollars. Credit card fraud is one of the most serious threats that modern businesses face. To effectively resist the illusion, it is necessary to first understand its mechanisms. Credit card thieves use a variety of techniques. The unauthorized use of another person's credit card for financial gain while both the card's rightful owner and the card issuer are unaware of the transaction is a common definition of credit card fraud. Card fraud typically begins with the loss or theft of the card itself or of the card account number or other sensitive data associated with the account that must be made available to a merchant during a legal transaction. The data is recorded on a magnetic stripe on the back of the card in a machine-readable format, and the card number, which is frequently the Primary Account Number (PAN), is displayed on the front. Cardholder's name, Card number, Expiration Date, Verification/CVV code, and Card Type are required fields. Credit card fraud can also be committed in a variety of other ways. Con artists are very skilled and can act quickly. This article will assist in identifying Application Fraud, which occurs when a person knowingly provides inaccurate information when applying for a credit card using the standard approach. Lost or stolen credit cards account for a large share of credit card theft. Aside from the traditional techniques of credit card fraud, more sophisticated fraudsters use skimming and tampering to steal money from unsuspecting victims. The magnetic strip on the back of the card or the information stored on the smart chip on the card can be copied from one card to another, providing the recipient access to the required information.

Literature review

The book by Rimpal R. Popat and Jayesh Chaudhary discusses They conducted a survey regarding the detection of credit card fraud, focusing on its three primary subfields: bank fraud, business fraud, and insurance fraud. They have prioritized the two primary methods of processing credit card transactions: I remotely (card not present) and ii in person. They have focused on techniques such as Regression, classification, Logistic regression, Support vector machine, Neural network, Artificial Immune system, K-nearest Neighbor, Naive Bayes, Genetic Algorithm, Data mining, Decision Tree, etc. They give a conceptual framework for six data mining approaches (classification, clustering, prediction, outlier detection, Regression, and visualization). The Artificial Immune System (AIS), Bayesian Belief Network (BBN), Neural Network (NN), Logistic Regression (LR), Support Vector Machine (SVM), Tree, Self-Organizing Map (SOM), and Hybrid Methods were among the various statistical and computational techniques covered (HM). All of the above-mentioned machine learning techniques, they reasoned, can deliver a high detection rate of accuracy, and businesses are eager to discover fresh ways to boost their earnings and cut expenses. Machine learning could be a viable alternative if you're seeking for a solution.

Authors Mohamad Zamini

Aiming to use auto encoder-based clustering for unsupervised fraud detection they have employed the auto encoder, which is an auto associator neural network, to reduce the dimensionality, extract the relevant features, and boost the network's learning performance. They trained their auto encoder-based clustering with the following parameters using a European dataset of 284,807 transactions, of which 0.17 percent were fraudulent. Three Hundred Iterations Cluster count = 2 When starting to cluster, k-means++ is the initialization to use. Acceptable level of divergence = 0.001 the model's learning rate is 0.01%. Epoch count = 200 the activation function is denoted by the symbols elu and Relu. Their context- free model design yielded a training loss of 0.024, a validation loss of 0.027, and a mean non-fraud data error of 75% less than the mean of reconstructive error, which was 25%. Regarding the model's predictions, the True positives equal 56,257, the false negatives equal 607, the False positives equal 18, the True negatives equal 80, and the best preferred equals (56,257 + 80 = 56,337). A total of 56,337 out of a possible 284,807 forecasts were accurate.

ShiyangXuan was presented, and a comparison was done between two random forests. Forests constructed using a combination of random trees (CART) and randomization (random). While both systems classify transactions as normal or abnormal, their basis classifications and performance are different, therefore they employ separate random forest strategies to train the behaviour aspects of each. Using data from a Chinese e-commerce firm,they tested both systems. where the percentage of fraudulent transactions in the subgroups ranges from one to ten. As a result, the CART-based random forest achieves a 96.7% accuracy, whereas the random-tree-based random forest achieves just 91.96 percent. Many issues, such as uneven data, have arisen because of the use of the B2C dataset. As a result, the algorithm can be enhanced.

These machine learning algorithms for detecting credit card fraud were proposed by author DejanVarmedja, who also conducted the research necessary to get these conclusions. Logistic Regression, Naive Bayes, Random Forest, and Multilayer Perceptron are some of the many machine learning techniques available. To achieve the best results, we employ an artificial neural network (ANN) called a multilayer perception, which consists of four hidden layers, a relu activation function (to prevent the usage of negative values), and an optimizer called Adam. Therefore, the Logistic regression accuracy score is 97.46%, with 56962 samples and 98 fraud transactions making up the data set. Accuracy scores of 99.23% and 99.96% are achieved by Naive Bayes and Random Forest, respectively, on the same dataset. The final ANN accuracy was 99.93%, and it was found that random forest provides the greatest outcome when it comes to detecting credit card fraud.

Authors Changjun Jiang proposed a new four-stage approach to fraud detection. The first stage involves using historical transaction data to divide transactions into clusters of similar behavior. The authors then developed a sliding window strategy to aggregate transactions. This algorithm is used to characterize a cardholder's behavioral pattern. Behavior patterns and responsibilities are finally sorted out and classified. Therefore, when compared to other methods, their approach, which combines Logistic Regression with raw data (RawLR), Random Forest with aggregation data (AggRF), and a feedback mechanism with aggregation data (AggRF +FB), achieves 80% accuracy.

.

Methodology

Keeping up with the Modeling and pattern of illicit transactions is getting harder as technology evolves. This labor-intensive method of detecting credit card fraud can now be automated, thanks to advancements in machine learning, artificial intelligence, and other related areas of information technology. In this research, the provided methods are applied to the problem of credit card fraud detection. Machine learning techniques like Logistic Regression, Decision Trees, Random Forest, Naive Bayes, SVM, and the K-Near classifier are compared to find the most effective one for

detecting fraudulent credit card purchases.

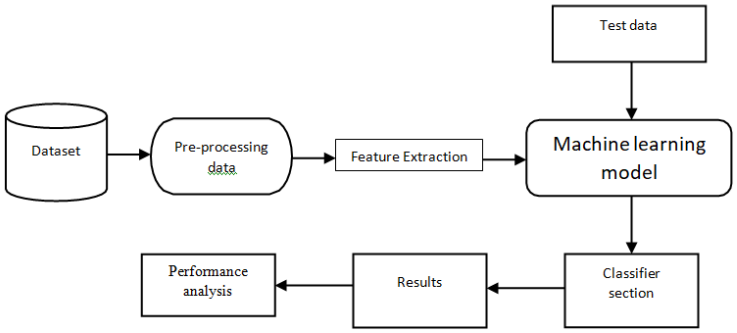

Fig. 1 proposed architecture

Simply said, it's predicated on a Machine Learning model that could boost the efficiency of specialized computer vision programmers. The proposed method is grounded in particular configurations of the Machine Learning model, as shown in Figure 1.

(HDP-MLT) utilizes an input dataset to generate a set of prediction models in a pipeline format. At a ratio of 80:20, it separates the dataset into training and testing data. After then, a number of ML models are cycled through in an iterative process to spot signs of transaction fraud. Various indicators are used to compare the effectiveness of the various models. Accuracy, precision, recall, and F1-score are all calculated using the model-specific confusion matrices. The method returns both the results of the fraud detection and performance statistics.

Performance Evaluation Metrics

In many ML-based situations, the confusion matrix serves as the foundation for creating metrics used to evaluate performance. Many works, including [2], [3], [7], and [10], investigate the usefulness of the confusion matrix. There are two examples where the predictions are accurate (TP and TN) and two examples where the forecasts are off (FP and FN).

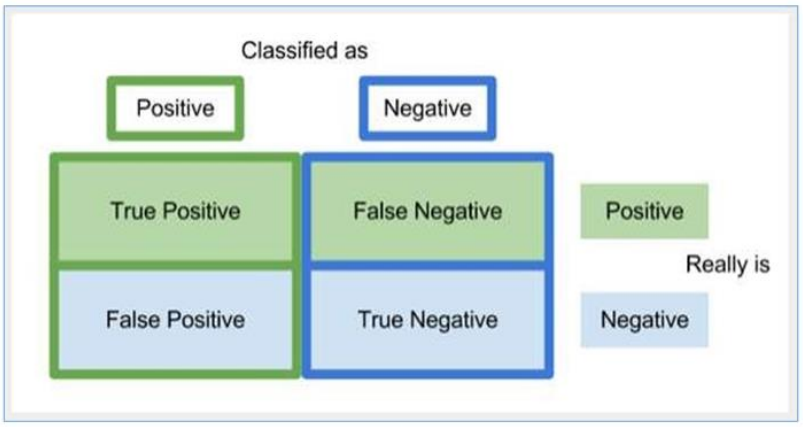

Figure 3: Confusion matrix model

Figure 3 shows how the performance metrics are derived by applying the confusion matrix toa variety of scenarios. The metrics for success are given in Equations (4)–(6). (7).

$$\text{Precision} = \frac{TP}{TP+FP} \quad (4)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (5)$$

$$\text{F1-measure} = 2 * \frac{(precision * recall)}{(precision+recall)} \quad (6)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

The values of these indicators range from 0 to 1, with 0 being the worst potential performance and 1 the best. When talking about values, higher means better.

Dataset description

This dataset represents hypothetical charges made to credit cards between January 1, 2019, and December 31, 2020, and may include both genuine and fraudulent charges. It protects thecredit card accounts of one thousand people who shop at a select group of 800 stores. Using Brandon Harris's Sparkov Data Generation | Github tool, we generated this. The time span of this simulation was from January 1, 2019, to December 31, 2020. Together, the files were

transformed into a universal standard.

Card numbers and other sensitive information were hashed because of a confidentiality agreement between the bank and the paper's authors. As a result of the discrepancy between the number of honest dealings and the number of scams, the overall data set had an extremelyasymmetrical distribution.

RESULTS AND DISCUSSION

Using AI for detecting fraudulent activity is a hot issue right now. An algorithm designed to detect credit card fraud looks for specific types of suspicious activity by comparing them to known fraud patterns. There are three distinct categories of machine learning, although the supervised and hybrid methods are best suited to detecting fraudulent activity. Here, we takea look at some of the latest developments in credit card fraud detection algorithms and compare them to some of the established methods of classification.

	Model	Train score	Test Score/Accuracy	Precision	Recall	F1-score	specificity
2	Decision Tree	94.78	94.92	94.013	94.139	94.076	95.500
3	Random Forest	90.22	89.68	96.216	79.041	86.787	97.667
5	K Nearest Neighbour	87.91	81.91	84.850	70.382	76.942	90.567
0	Logistic Regression	86.84	86.60	94.178	73.268	82.418	96.600
4	Naive Bayes	81.55	81.53	97.411	58.481	73.085	98.833
1	Support Vector Machines	47.79	45.49	43.992	99.334	60.979	5.067

Fig 4. Comparison table for performance of models

As shown in Figure 3, ML is utilized for the identification of credit card fraud. The accuracy of various models is compared to that of other employed models. There is significant performance improvement when recommended method is applied. Each method exhibited an increased F1-score compared to the models utilized in conventional methods.

Confusion matrix of proposed algorithms

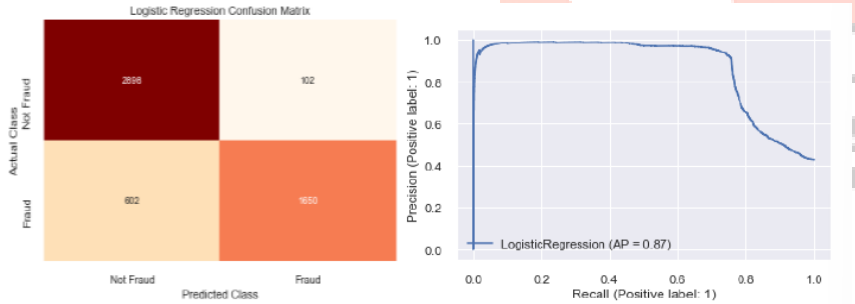

Fig.5.Confusion matrix of Logistic regression

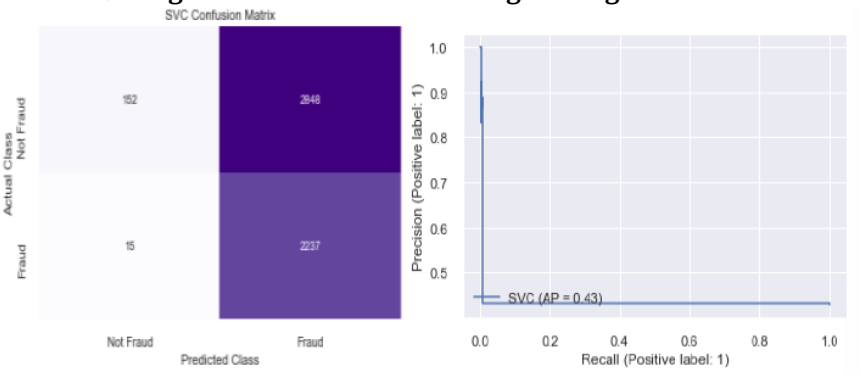

Fig.6.Confusion matrix of SVC

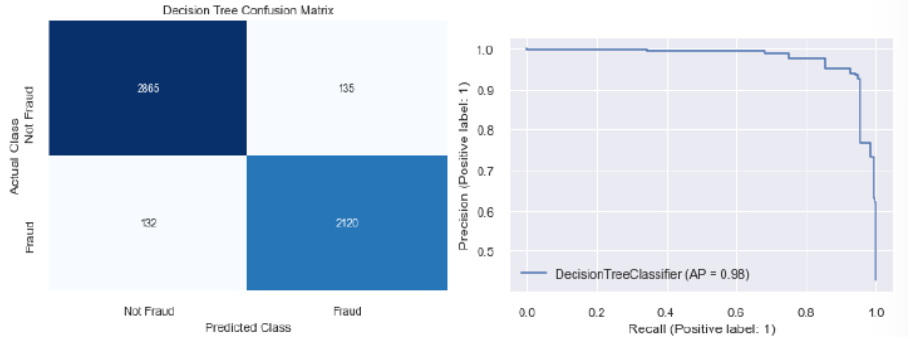

Fig.7.Confusion matrix of Decision Tree

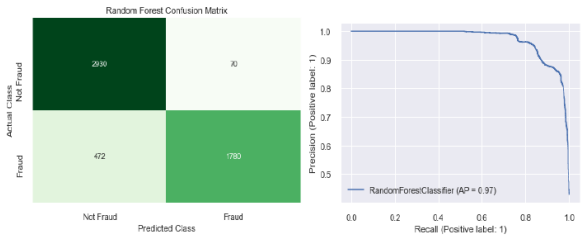

Fig.8.Confusion matrix of Random Forest

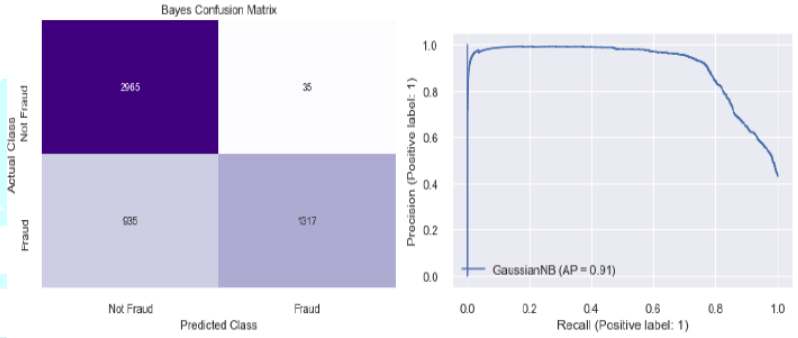

Fig.9.Confusion matrix of native bayes

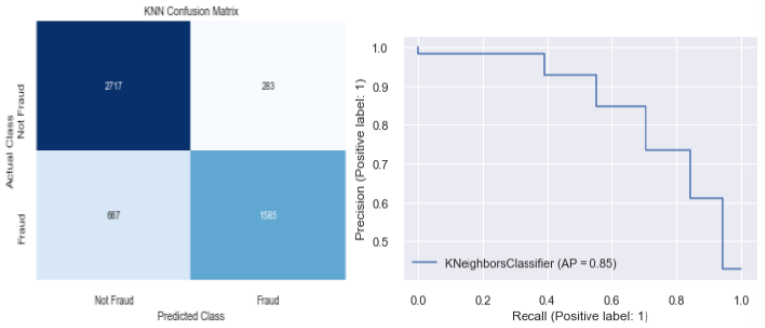

Fig.10.Confusion matrix of KNN

A confusion matrix, which is depicted in the graphics above, is a table used to characterize the performance of a classification method. A confusion matrix is a graphical representation and summary of the performance of a classification algorithm. The confusion matrix consists of four fundamental properties (numbers) utilized to define the classifier's measuring metrics. These four digits represent:

TP (True Positive): TP indicates the number of transactions who have been correctly identified as having malignant lymph nodes, indicating that they have the illness.

TN (True Negative): TN indicates the number of accurately identified healthy patients. FP (False Positive): FP is the number of incorrectly transactions who are truly healthy. FP isreferred to as a Type I error.

4. FN (False Negative): FN shows the number of patients misclassified as healthy althoughactually they are suffering from the disease. FN is also known as a Type II mistake.

Performance measures of an algorithm are accuracy, precision, recall, and F1 score, whichare determined on the basis of the above-stated TP, TN, FP, and FN.

Conclusion

We have achieved an accuracy score of 94.79 by using our improved Decision tree algorithm to tackle the issue of credit card fraud detection. In comparison to findings acquired using the existing modules, the suggested module may be applied to a larger dataset and generates results that are more trustworthy. The Random forest method performs

better with additional training data, but it still falls short in testing and real application. Additionally, it might be beneficial to apply more pre-processing techniques.

.

References

[1] ShiyangXuan, Guanjun Liu, Zhenchuan Li, LutaoZheng, Shuo Wang, Changjun Jiang, Random Forest for Credit Card Fraud Detection,2018 (IEEE).

[2]. DejanVarmedja, MirjanaKaranovic, SrdjanSladojevic, Marko Arsenovic, and AndrasAnderla,Credit Card Fraud Detection - Machine Learning methods, Publish in:18th International Symposium INFOTEH-JAHORINA, 20-22 March 2019 (IEEE).

[3]. ShantanuRajora, Dong-Lin Li, ChandanJha, NehaBharill, Om Prakash Patel, Sudhanshu Joshi, Deepak Puthal, MukeshPrasad,A Comparative Study of Machine Learning Techniques for Credit Card Fraud Detection Based on Time Variance,2018 (IEEE).

[4]. Rimpal R. Popat and JayeshChaudhary,A Survey on Credit Card Fraud Detection using Machine Learning, 2018 (IEEE), pp. 1120 - 1125

[5]. Changjun Jiang, Jiahui Song, Guanjun Liu, Member, IEEE, LutaoZheng, and WenjingLuan,Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism,2018 (IEEE), pp. 2327-4662.

[6]. Shail Machine, Emad A. Mohamad, BehrouzFar,Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study, 2018(IEEE) computer society, pp.122-125.

[7]. SamanehSorournejad, Zahra Zojaji, Reza EbrahimiAtani, Amir Hassan Monadjemi,A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective.

[8]. Kuldeep Randhawa, Chu Kiong Loo, ManjeevanSeera, Chee Peng Lim, Ashoke K. Nandi,Credit Card Fraud Detection Using AdaBoost and Majority Voting, Published in: IEEE Access on 15 February 2018, vol. no.6, pp. 14277 – 14283.

[9]. N. Sivakumar, Dr.R. Balasubramanian, Fraud Detection in Credit Card Transactions: Classification, Risks and Prevention Techniques, Published in (IJCSIT) International Journal of Computer Science and Information Technologies, vol no. 6 (2), 2015, pp. 1379-1386.

[10]. SaiKiran, Jyoti Guru, Rishabh Kumar, Naveen Kumar, Deepak Katariya, MaheshwarSharma,Credit card fraud detection using Naïve Bayes model based and KNN classifier, Published in: International Journal of Advance Research, Ideas and Innovations in Technology, Issue no. 3, vol.no. 4, 2018, pp.44-47.

[11]. Rishi Banerjee, Gabriela Boural, Steven Chen, MehalKashyap, Sonia Purohit, Jacob Battipagali,Comparative Analysis of Machine Learning Algorithms through Credit Card Fraud Detection, 2018.

[12]. KaithekuzhicalLeenaKurien and Dr.AjeetChikkamannur,Detection and prediction ofcredit card fraud transactions using machine learning, Published in: International Journal of engineering science & research technology (IJESRT), 2019, pp. 199-208.

[13]. Md.Akster Hossain and Mohammed NazimUddin,A Differentiate Analysis for Credit Card Fraud Detection, Published in: Int. Conf. on Innovations in Science, Engineering and Technology (ICISET), 27- 28 October 2018 (IEEE), pp. 328-333.

[14]. Suresh K Shirgave, Chetan J. Awati, Rashmi More, Sonam S. Patil,A Review on Credit Card Fraud Detection Using Machine Learning, Published in International journal of Science and Technology Research, vol.no.8, Issue no. 10, October 2019, pp. 1217-1220.

[15]. S P Mani raj, Aditya Saini, Swarna Deep Sarkar ShadabAhmed,Credit Card Fraud Detection using Machine Learning and Data Science, Published in: International Journal of Engineering Research & Technology (IJERT), vol.no.8, Issue no. 09, September 2019, pp.110-115.