International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

System for Detecting Credit Card Fraud Using Intelligent Retrieval and Machine Learning

Dr. Pritam R. Patil[1], Mr. Pranav P. Joshi[2], Dr. N N Bharkad[3],
Mr. Amol V Suryawanshi[4], Dr.P A Kadam[5]
[1]Asst. Prof., Institute of Technology & Management, Nanded, Maharashtra, India
[2]Asst. Prof., Institute of Technology & Management, Nanded, Maharashtra, India
[3]Asst. Prof., Institute of Technology & Management, Nanded, Maharashtra, India
[4]Asst. Prof., Institute of Technology & Management, Nanded, Maharashtra, India
[5]Asst. Prof., Institute of Technology & Management, Nanded, Maharashtra, India
E-Mail: pritam.itm@gmail.com

Abstract:

Credit card fraud detection is currently taking place on a huge scale all around the world. This issue persists despite a significant increase in online transactions and use of e-banking platforms. It is critical that credit card firms be able to detect fraudulent credit card transactions so that users are not overcharged. Such issues can be addressed using Intelligent Retrieval and Machine Learning.
Credit card fraud usually occurs when the card is stolen and used for unauthorised purposes, or when the fraudster is able to extract the credit card information for their own use. To detect such fraudulent acts, the credit card fraud detection system was established. The project's goal is to concentrate mostly on machine learning methods. The Random Forest, Support Vector Machines, XG Boost, Logistic Regression are all employed.
The above suggested system's performance will be evaluated using sensitivity, specificity, accuracy, and error rate. By comparing all three methods, the optimal Algorithm will be discovered.

Keywords: XG Boost, Logistic Regression, Random Forest

INTRODUCTION

The term "credit card fraud" refers to a broad range of fraudulent and theft acts that occur when a person uses their credit card to transfer payments. This could serve two purposes, either to take money out of an account without authorization or to make purchases without paying for them directly. Identity theft is also a result of credit card fraud. Credit card fraud is the crime that most people identify with identity theft, despite the fact that it is somewhat of an economic crime.

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

Approximately 10 million, or one out of every 1300 transactions, were made in 2000 out of the 13 billion transactions that are made annually proved to be a hoax. Additionally, 5 out of 10,000 monthly active accounts, or 0.05% of them, were fraudulent.
Currently, one-twelfth of one percent of all transactions are protected by fraud detection systems, but billions of dollars are still lost as a result. Today's business premises are extremely vulnerable to credit card fraud. Therefore, it is crucial to comprehend the methods used to carry out a fraud in order to effectively resist it. There are undoubtedly a plethora of methods that credit card thieves use to conduct fraud. To put it simply, credit card fraud is described as "when a specific person uses another person's credit card for independent purposes without disclosing the card's ownership or issuer of any activity on the card." Card fraud can begin with the physical card being stolen or with the theft of sensitive account information, such as the card account number or other information that must be made available to a merchant in order for a transaction to be approved. As a result, identifying credit card fraud is typically challenging. Machine learning is considered to be one of the most effective methods for detecting fraud. Regression and classification techniques are used to predict credit card fraud. A number of learning algorithms have been investigated for credit fraud detection cards that include Random Forest, Logistic Regression Support Vector Machines, and Neural Networks. The above algorithms' performance is verified by this project. Their capacity to indicate if a transaction was fraudulent or authorised will determine how well they compare. Accuracy, specificity, and precision of performance measures are used to facilitate the comparison.

LITERATURE SURVEY

Algorithms from both machine learning and deep learning are used in credit card fraud research. This section describes the work that was done using two distinct methodologies that are easily accessible for detecting fraud, and the methods that can be used to deal with data that is not balanced. Some methods are available to deal with the unbalanced data. They are sampling techniques (b), classification techniques (a), and resemblance approaches (c). The following Machine Learning algorithms are used to detect credit fraud: K-nearest neighbour, decision trees, gradient boosting, and logistic regression, support vector machine (SVM), etc. Yashvi Jain, Namrata Tiwari, Shripriya Dubey, and Sarika Jain investigated a number of methods [I] in 2019 for the identification of credit card fraud, including support vector machines (SVM), Bayesian networks, decision trees, K-Nearest Neighbours (KNN) fuzzy logic system, hidden markov model, artificial neural networks (ANN), and fuzzy logic systems. They note in their research that the SVM, decision trees, and knearest neighbour algorithms all provide medium levels of accuracy. Out of all the algorithms, the ones with the lowest accuracy are Fuzzy Logic and Logistic Regression. High detention rates are provided by KNN, fuzzy systems, naive bayes, and neural networks. At the medium level, the SVM, decision trees, and logistic regression all provide a high detection rate. The ANN and the Naïve Bayesian Networks are the two methods that outperform in every regard. It is highly costly to train these. There is a significant flaw in every algorithm. The disadvantage is that different contexts get different results from these algorithms. With one kind of dataset, they provide better findings; with another type of dataset, they produce poorer results. Small datasets do exceptionally well

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

with algorithms like KNN and SVM, whereas raw and un sampled data work well with methods like logistic regression and fuzzy logic systems.

PROPOSED SOLUTION

This project's primary goal is to use algorithms such the Random Forest, Decision Tree, LGBM, and Nearest Neighbours algorithms to classify the transactions in the dataset that contain both fraud and non-fraud transactions. The optimal algorithm for identifying credit card fraud transactions is then determined by comparing these three algorithms. The process flow for detecting credit fraud involves dividing the data, training the model, deploying the model, and establishing evaluation standards. This model uses the Kaggle credit card fraud dataset, which needs to undergo pre-processing. We now need to divide the data into training and testing sets in order to create the model. The Random Forest and LGBM models are trained using the training set. Next, we create both models. Lastly, the models' accuracy, precision, recall, and F1-score are computed. At last, a more precise comparison of credit card fraud transactions.

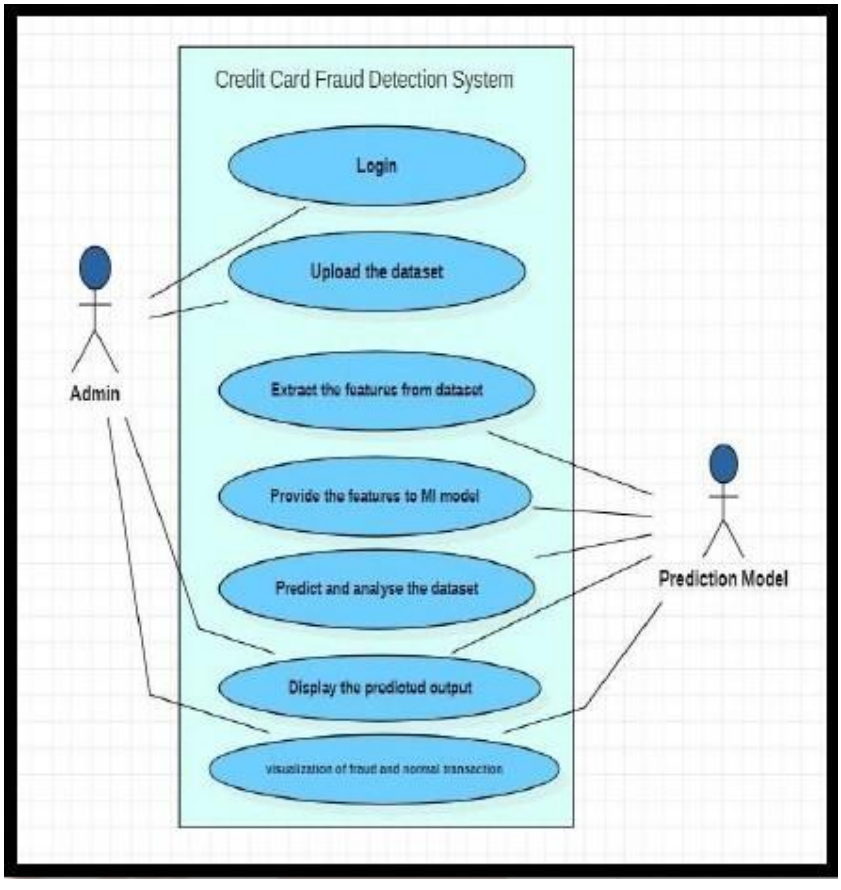

Fig. 1 Uml Diagram

To begin with examine the dataset. To balance the data set, random sampling is applied. Next the dataset should be divided into two sections: the train dataset and the test dataset. Move on the suggested models employ feature selection. Metrics for accuracy and performance have

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

been computed to determine the effectiveness of various algorithms. Next, find the optimal algorithm for the specified dataset by evaluating its efficiency.

- **Algorithm to be used – Random Forest, Decision tree, LGBM and the Nearest neighbours.**

METHODOLOGY

The process of detecting credit card fraud looks like this. The dataset can be uploaded in one or more files, and the algorithm will read them all. To balance the data set, random sampling is applied. Next the dataset should be divided into two sections: the train dataset and the test dataset. Move on The suggested models employ feature selection. The effectiveness of various algorithms has been determined by calculating accuracy and performance indicators. Next, find the optimal algorithm for the specified dataset by evaluating its efficiency.

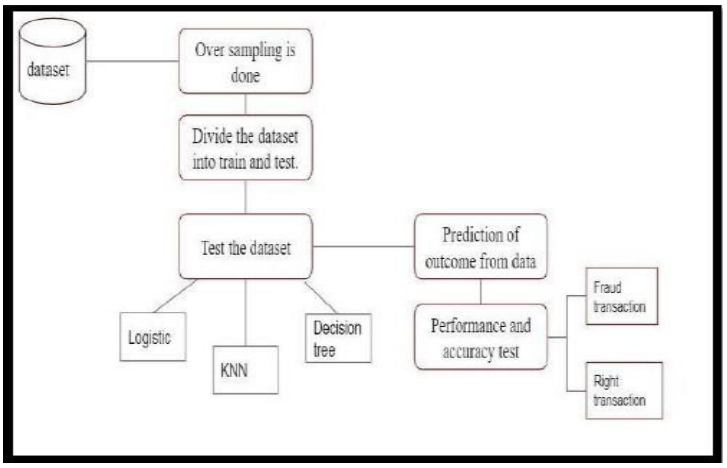

Fig 2 Methodology

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

Step 1: Import the dataset
Step 2: Convert the data into data frames format
Step3: Do random oversampling using ROSE package
Step4: Decide the amount of data for training data and testing data
Step5: Give 70% data for training and remaining data for testing.
Step6: Assign train dataset to the models
Step7: Choose the algorithm among 3 different algorithms and create the model
Step8: Make predictions for test dataset for each algorithm
Step9: Calculate accuracy for each algorithm
Step10: Apply confusion matrix for each variable
Step11: Compare the algorithms for all the variables and find out the best algorithm.

Table 1. Algorithm steps for finding the Best algorithm

Results

```
data.head()
```

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...	-0.018307	0.277838	-0.110474
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...	-0.225775	-0.638672	0.101288
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	...	0.247998	0.771679	0.909412
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...	-0.108300	0.005274	-0.190321
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	...	-0.009431	0.798278	-0.137458

5 rows × 31 columns

```
len(data)
```

```
284807
```

Fig 3.Data set

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

```
from sklearn.metrics import accuracy_score, f1_score, precision_score, recall_score, plot_roc_curve
acc = accuracy_score(y_test, y_pred)
prec = precision_score(y_test, y_pred)
rec = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
print('accuracy:%0.4f'%acc,'\tprecision:%0.4f'%prec,'\trecall:%0.4f'%rec,'\tF1-score:%0.4f'%f1)
```

```
accuracy:0.9995         precision:0.9417        recall:0.7687   F1-score:0.8464
```

```
### Store results in dataframe for comparing various Models
results_testset = pd.DataFrame([['RandomForest', acc, 1-rec, rec, prec, f1]],
             columns = ['Model', 'Accuracy', 'FalseNegRate', 'Recall', 'Precision', 'F1 Score'])
results_testset
```

	Model	Accuracy	FalseNegRate	Recall	Precision	F1 Score
0	RandomForest	0.99952	0.231293	0.768707	0.941667	0.846442

Fig4.Algorithm

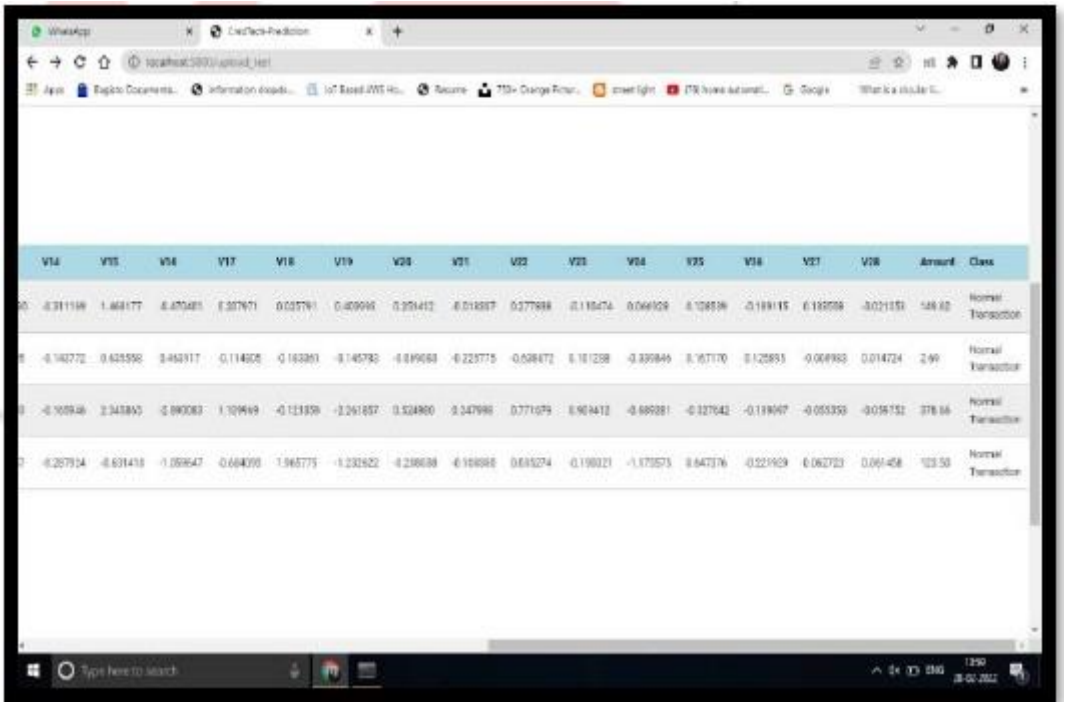

Fig 5.Output

Conclusion

In this study, fraud in the credit card system was detected using machine learning techniques such as random forest, SVM, and logistic regression. Sensitivity, specificity, accuracy, and error rate are used to assess how well the suggested system performs. Through comparison of all three methods, the optimal algorithm is identified.

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

References

1. John Richard D. Kho, Larry A. Vea published by Proc. of the 2017 IEEE Region 10 Conference (TENCON) in "Credit Card Fraud Detection Based on Transaction Behaviour", Malaysia, November 5-8, 2017

2. Clifton Phua, Vincent Lee, Kate Smith & Ross Gayler, A Comprehensive Survey of Data Mining-based Fraud Detection Research published by School of Business Systems, Faculty of Information Technology, Monash University, Wellington Road, Clayton, Victoria 3800, Australia

3. Hisar HCE, Sonepat published by International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 3, March 2014

4. Wen-Fang YU and Na Wang published by 2009 International Joint Conference on Artificial Intelligence Research 'Credit Card Fraud Detection Model Based on Distance Sum'.

5.Massimiliano Zanin, Miguel Romance, Regino Criado, and SantiagoMoral published by Hindawi Complexity 'Credit Card Fraud Detection through Parenclitic Network Analysis' Volume 2018, Article ID 5764370, 9 pages

6. Iwasokun GB, Omomule TG, Akinyede RO. Encryption and tokenization-based system for credit card information security. Int J Cyber Sec Digital Forensics. 2018;7(3):283–93.

7. Sriram Sasank JVV, Sahith GR, Abhinav K, Belwal M. Credit card fraud detection using various classification and sampling techniques: a comparative study. In: IEEE, 2019. p. 1713–1718.

8. Ojugo AA, Nwankwo O. Spectral-cluster solution for credit-card fraud detection using a genetic algorithm trained modular deep learning neural network. JINAV J Inf Vis. 2021;2:15–24. https://doi.org/10.35877/454RI.jinav274.

9. Darwish SM. An intelligent credit card fraud detection approach based on semantic fusion of two classifiers. Soft Comput. 2019;24:1243–53. https://doi.org/10.1007/s00500-019-03958-9.

10. Li C, Ding N, Dong H, Zhai Y. Application of credit card fraud detection based on CS-SVM. Int J Mach Learn Comput 2021;11(1).

11. Olowookere TA, Adewale OS. A framework for detecting credit card fraud with cost-sensitive meta-learning ensemble approach. Sci Afr. 2020;8:e00464. https://doi.org/10.1016/j.sciaf.2020.e00464.

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

12. Itoo F, Meenakshi SS. Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. Int J Inf Technol. 2020; 13:1503–11. https://doi.org/10.1007/s41870-020-00430-y.