Project : Credit Card Fraud Detection

[Semester – I of I Year M.Sc.(Cyber Security) 2023-24]

Submitted By :

Group ID : 4

Enroll No.	Name of Students
2057	Yash Gajera
2056	Astha Dhorajiya
2053	Pritesh Patel

Date of Submission : 06-12-2023

Submitted To : Dr. Ravirajsinh Vaghela

National Forensic Science University

Introduction

Now, as with each passing year more and more countries are going cashless and the dependency on online payment methods are increasing, many complicated investigation systems are being developed and used so as to identify obscure patterns and the relationships among large informational indexes which were earlier impossible to detect. This comes under a new term called Information Mining. The tools which are used in information mining are
1) Machine learning methods
2) Factual models
3) Mathematical calculations

Credit card fraud is a major concern for both consumers and financial institutions. Fraudulent transactions can lead to financial losses and damage to the reputation of financial institutions. Machine learning techniques have been used extensively to detect fraudulent transactions. In this project, we use logistic regression to classify transactions as either legitimate or fraudulent based on their features.

Dataset Description

The dataset contains transactions made by a cardholder in a duration in 2 days i.e., two days in the month of September 2013. Where there are total 284,807 transactions among which there are 492 i.e., 0.172% transactions are fraudulent transactions. This dataset is highly unbalanced. Since providing transaction details of a customer is considered to issue related to confidentiality, therefore most of the features in the dataset are transformed using principal component analysis (PCA). V1, V2, V3,..., V28 are PCA applied features and rest i.e., 'time', 'amount' and 'class' are non-PCA applied features, as shown in table

Table : Attributes of European dataset

S. No.	Feature	Description
1.	Time	Time in seconds to specify the elapses between the current transaction and first transaction.
2.	Amount	Transaction amount
3.	Class	0 - not fraud 1 – fraud

Reference :- https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

Preprocessing

Before training the model, we first separate the legitimate and fraudulent transactions. Since the data is imbalanced, with significantly more legitimate transactions than fraudulent transactions, we undersample the legitimate transactions to balance the classes. We then split the data into training and testing sets using the train_test_split () function.

In this dataset we done the following data preprocessing step:
1- Acquire the dataset
2- Import all the crucial libraries
3- Import the dataset
4- Identifying and handling the missing values
5- Encoding the categorical data
6- Splitting the dataset
7- Feature scaling

Model

Logistic Regression

We use logistic regression to classify transactions as either legitimate or fraudulent based on their features. Logistic regression is a widely used classification algorithm that models the probability of an event occurring based on input features. The logistic regression model is trained on the training data using the LogisticRegression () function from scikit-learn. The trained model is then used to predict the target variable for the testing data.

Decision Tree Classifier

The aim of the Decision Tree model is to build a small decision tree with high precision. Based on credit card fraud detection, the decision tree has two stages. The initial step is to build a decision tree using the training data provided, and the later step is to use decision rules to classify incoming transactions.The decision tree's input data is labelled with class labels, such as legitimate or fraudulent.

Random Forest Classifier

The random forest algorithm was chosen because it acts as a binary classifier, which makes it suitable for credit card fraud detection, as soon as any transaction is classified as one of two classes (0 or 1) fraud or not a fraud.Even for large data sets with many features and data instances training is extremely fast in random forest.The Random Forest algorithm has been found to provide a good estimate of the generalization error and to be resistant to over fitting.

SVM

SVM is a binary classification, hence the transactions are labelled either as fraudulent, or legitimate. This helps us to identify abnormal behaviour of user i.e. Fraud User. It uses regression technique.The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

Output of Training Dataset

```
Training Dataset
```

:

	Name	DecisionTreeClassificaton	Randomforestclassifier	LogisticRegression	SVM
0	Accuracy	1.0	1.0	0.926302	0.898348
1	Precision	1.0	1.0	0.890863	0.817259
2	Recall	1.0	1.0	0.959016	0.975758
3	F1-Score	1.0	1.0	0.923684	0.889503
4	Roc_Auc_score	1.0	1.0	0.928439	0.909104

Output of Testing Dataset

```
Testing Dataset
```

Out[34]:

	Name	DecisionTreeClassificaton	Randomforestclassifier	LogisticRegression	SVM
0	Accuracy	0.918782	0.913706	0.893401	0.888325
1	Precision	0.918367	0.846939	0.846939	0.795918
2	Recall	0.918367	0.976471	0.932584	0.975000
3	F1-Score	0.918367	0.907104	0.887701	0.876404
4	Roc_Auc_score	0.918780	0.921271	0.896848	0.902030

Conclusion

Credit card fraud is without a doubt an act of criminal dishonesty. This article has listed out the most common methods of fraud along with their detection methods and reviewed recent findings in this field. This paper has also explained in detail, how machine learning can be applied to get better results in fraud detection along with the algorithm, pseudocode, explanation its implementation and experimentation results.

While the algorithm does reach over 99.6% accuracy, its precision remains only at 28% when a tenth of the data set is taken into consideration. However, when the entire dataset is fed into the algorithm, the precision rises to 33%. This high percentage of accuracy is to be expected due to the huge imbalance between the number of valid and number of genuine transactions.

Since the entire dataset consists of only two days' transaction records, its only a fraction of data that can be made available if this project were to be used on a commercial scale. Being based on machine learning algorithms, the program will only increase its efficiency over time as more data is put into it.

Future Enhancements

While we couldn't reach out goal of 100% accuracy in fraud detection, we did end up creating a system that can, with enough time and data, get very close to that goal. As with any such project, there is some room for improvement here.

The very nature of this project allows for multiple algorithms to be integrated together as modules and their results can be combined to increase the accuracy of the final result.

This model can further be improved with the addition of more algorithms into it. However, the output of these algorithms needs to be in the same format as the others. Once that condition is satisfied, the modules are easy to add as done in the code. This provides a great degree of modularity and versatility to the project.

More room for improvement can be found in the dataset. As demonstrated before, the precision of the algorithms increases when the size of dataset is increased. Hence, more data will surely make the model more accurate in detecting frauds and reduce the number of false positives.

References

1. "Credit Card Fraud Detection Based on Transaction Behaviour -by John Richard D. Kho, Larry A. Vea" published by Proc. of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 2017
2. CLIFTON PHUA1, VINCENT LEE1, KATE SMITH1 & ROSS GAYLER2 " 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. "Survey Paper on Credit Card Fraud Detection by Suman" , Research Scholar, GJUS&T Hisar HCE, Sonepat published by International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 3, March 2014
4. "Research on Credit Card Fraud Detection Model Based on Distance Sum – by Wen-Fang YU and Na Wang" published by 2009 International Joint Conference on Artificial Intelligence
5. "Credit Card Fraud Detection through Parenclitic Network Analysis-By Massimiliano Zanin, Miguel Romance, Regino Criado, and SantiagoMoral" published by Hindawi Complexity Volume 2018, Article ID 5764370, 9 pages