Research Paper E-ISSN NO : 2455-295X | VOLUME : 9 | ISSUE : 6 | SPECIAL ISSUE JUNE-2023

FRAUD DETECTION IN CREDIT CARD TRANSACTION USING MACHINE LEARNING

DR. A. S. SHANTHI	PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, TAMILNADU COLLEGE OF ENGINEERING, COIMBATORE, INDIA- 641 659.
MRS. T.RAMYA	ASSISTANT PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, TAMILNADU COLLEGE OF ENGINEERING, COIMBATORE, INDIA- 641 659.
V.SIVA HARIHARAN	STUDENT, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, TAMILNADU COLLEGE OF ENGINEERING, COIMBATORE, INDIA- 641 659.
C.RAJASEKAR	STUDENT, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, TAMILNADU COLLEGE OF ENGINEERING, COIMBATORE, INDIA- 641 659.
G.DENEESH	STUDENT, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, TAMILNADU COLLEGE OF ENGINEERING, COIMBATORE, INDIA- 641 659.
K. DEEPA	STUDENT, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, TAMILNADU COLLEGE OF ENGINEERING, COIMBATORE, INDIA- 641 659.

ABSTRACT:

Due to increase of fraud which results in loss of money across the globe, several methodologies and techniques developed for detecting fraud detection involves analyse the activities of users on order to understand the malicious performance of users. Malicious behaviors are a broad term including delinquency, fraud, intrusion, and account defaulting. The number of online transitions has grown in large quantities a credit card transaction holds a huge share of these transactions. Therefore, bank and financial institutions offer fraud detection in credit card transaction application much value and demand. Fraudulent transaction can occur in various ways and can be but into different categories. This project focus on fraud occasions in real time-world transitions. This project presents a survey of current techniques used in fraud detection in credit card transaction and evaluates a new hybrid approach to identify fraud detection. In the project work, we analyze fraud detection using machine learning algorithm namely logistic regression and decision tree. Towards make the learning process resourceful, we use principal component analysis for feature selection.

KEYWORDS:

DETECTION, REGRESSION, DECISION TREE, PCA, FRAUD, TRANSITIONS.

INTRODUCTION

Fraud has been increasing significantly with the progression of state-of-art technology and worldwide communication. The credit card transaction information is confidential, the bank and the other financial enterprises to survive in such competing industry. Fraud can be avoided in two main ways: one is prevention and another one is detection. Prevention avoids any attacks from fraudsters by acting as a layer of protection. Recognition happens once the impediment has already failed. Consequently, detection helps in identify and alert as soon as a fraudulent transaction is individual trigger. In a real-world online payment the massive stream of payment requests is quickly scanned by automatic tools that determine which transactions to authorize. Classifiers are typically employed to analyze all the authorized transactions and alert the most suspicious ones. Alerts are then inspected by professional investigators that contact the card holders ¬¬¬¬to determine the true nature of each alerted transaction. By doing this, investigators provide a feedback to the system in the form of labelled transitions, which can be used to train or update the classifier, in order to preserve the fraud-detection performance over time. The vast majority of transactions cannot be verified by investigators time and cost constraints. These transactions remain unlabeled until customers discover and report frauds, or until a sufficient amount of time has elapsed such that none distributed transactions are considered genuine. A cash transaction log mining is one of the most significant fields in the area of machine learning. There have been a large number of machine learning algorithms rooted in these fields to perform different data analysis tasks. Apply principal component analysis algorithm to find best features, and then apply decision tree classifier for cash transaction fraud prediction.

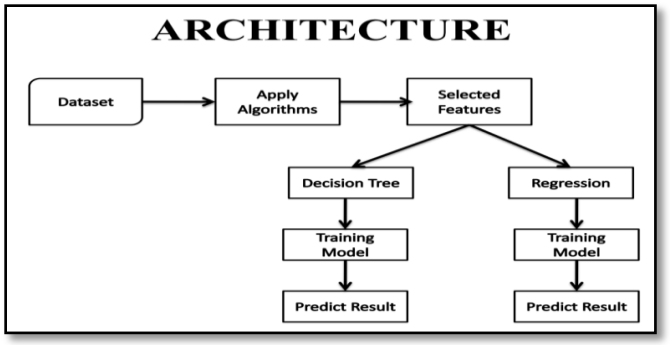

FIG 1. MODULES FRAMEWORK

A. DATA DESCRIPTION

The dataset was created combining two data sources; the fraud transactions log file and all transactions log file. The fraud transactions log file holds all the online credit card fraud occurrences while all transactions log file holds all transactions storied by the corresponding bank within a specified time period. Due to the confidential disclosure agreement made between the bank and the authors of the project, some of the sensitive attributes such as card number were hashed. When evaluating the combined dataset, the shape of the data was much skewed due to the imbalanced number of rightful transactions and fraudulent occurrences. The case with the fraud cases had 200 records while the operation log file had 917781 records. Attributes of the two data sources are as follows.

Field name	Description
CARD_NO	Credit cardholders hashed card number
DATE	Date of the transaction
TIME	Time of transaction
TRANSACTION_AMOUNT	Transaction amount
MERCHANT_NAME	Merchant name relevant to the transaction
MERCHANT_CITY	Registered city of the merchant
MERCHANT_COUNTRY	Registered country of the merchant
RESPONSE_CODE	ISO Response code related for the transaction
MERCHANT_CATEGORY_CODE	Category code of the Merchant
APPROVED/NOT APPROVED	Status of the transaction

TABLE 1: THE GENUINE TRANSACTIONS LOG

Field Name	Description
CARD_NO	Credit card holder's hashed card number
DATE	Date of the transaction
TIME	Time of transaction
SEQ	A unique sequence number which was given for frauds
NATURE OF THE FRAUD	Nature of the frauds as card present or not present
MCC	Category code of the merchant
AMOUNT OF FRAUD	Transaction amount
REVERSAL	Bank related field

TABLE 2: THE FRAUDULENT TRANSACTIONS LOG

B. DATA PREPARATION

Collected raw data were first divided into4 data sets according to its fraud pattern. This process was done with the information gained by the bank. The four datasets are,

1. Transactions with Risky Merchant Category code (MCC)
2. Transactions larger than $100.
3. Transactions with risky ISO Response code.
4. Transactions with unknown web addresses.

Those 4 datasets were used in two different ways.

1. By transforming raw data into a numerical form. (type A)
2. By preparing raw data categorically without making any transformation.(type B)

Type A was applied to datasets 1, 2, 3 and type B was applied to data set number 4. In data preparation the data are cleaned, transformed, integrated and reduced. First 3 sets of data were subjected to all above-mentioned steps to prepare them numerically. To prepare the data categorically all the steps except for data transformation were applied. The basic steps which were involved in the type A are described below.

DATA CLEANING – filling in missing values is an important task in the data cleaning process. There are many ways to overcome this issue, such as ignore the whole tuple, but most of them are likely to bias the data. Since the source file which contained genuine transaction did not contain records with missing values, filling them was no more an issue. Tuples with meaningless value were removed from the files as well as they do not bias the data. Additionally, following changes such as removing unnecessary columns, separating the data time column into two.

DATA INTEGRATION – before the data were subjected to further change the two sources were integrated together since fraudulent and genuine record files were in two separate files. I show how the mapping process was done.

Genuine transactions	Fraudulent transaction
Card number	Card number
Date and time	Date and time
MCC	MCC
Transaction amount	Transaction amount

TABLE 3: DATA MAPPING

Data transaction – here, all the categorical data were consolidate into an understandable numerical format. The transactional dataset contains several data types with several ranges. Therefore, data transformation comprises of data normalization. Data normalization scales the attribute data to fall in a small numeric range.

Data reduction – the strategy used for this is dimension reduction. We must prevent the risk of learning wrong patterns of data and the selected features should eliminate the irrelevant aspects and qualities of the fraud domain. The principal component analysis which is well-known that PCA is a popular transform method resolves the feature selection issue from the perspective selection successfully by finding the suitable number of principal components. In type B, data cleaning and data integration were involved as same as in type A. Then those data were taken to the next step of process.

C. RE-SAMPLING TECHNIQUES

The two data sources were characterized by a highly imbalanced distribution of examples among the classes. Fraudulent transactions contained a much smaller number of examples than the genuine transactions. To overcome this, we conducted under-sampling and over-sampling by reducing the majority occurrences and by raising the minority occurrences respectively. For over-sampling, Synthetic Minority Oversampling techniques (SMOTE) and for under-sampling, Condensed Nearest Neighbour (CNN) and Random Under-Sampling(RUS) were used. The minority class is over-sampled by producing "Synthetic" examples in SMOTE method. Out of RUS and CNN, RUS is a non-heuristic method to eliminate random majority class examples. We must prevent the risk of learning wrong patterns of data and the selected feature should eliminate the irrelevant aspects and qualities of the fraud domain. Then the data, to which the cross-validation was applied, were resampled by the above-mentioned re-sampling techniques.

D. MODELLING AND TESTING

Our project analyses four different fraud patterns. For analyzing each pattern, we have a few numbers of techniques were used in the data analysis. Four machine learning algorithms were prioritized in our analysis with the help of the literature. They are Support Vector Machine, Naive Bayes, K-Nearest Neighbor and Logistic Regression. We applied those selected Supervised Learning classifiers to our re-sampled data. When selecting machine learning models which can be capture each fraud, the accuracy and performance of each model were taken into consideration.

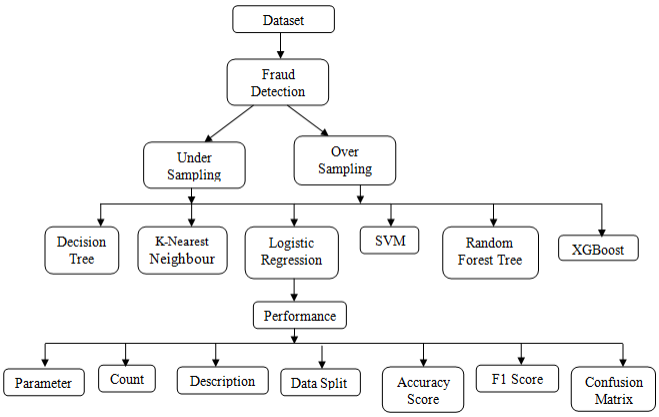

FIG 2: MODELLING AND TESTING

E. FRAUD DETECTION FOR REAL TIME:

Fraud detection has been done by taking already happen dealings in bulkiness and applying machine learning models on them. Since the results can be seen after weeks or months, tracking down of detected frauds was found particularly difficult, and there have been many cases where the fraudsters were able to commit many more fraudulent purchase earlier than being exposed. Fraud detection for real time is the carrying out of fraud detection models the second an online acquisition is taken place. That way our system is able of detecting frauds real-time.

F. FRAUD DETECTION

Real time detection of credit card fraud can be declared as one of the main contributions of this project. All the components are involved in fraud detection at the same time. The predict results and other main data of the machine learning models. The user can interact with the fraud detection system it shows the real time transactions, alerts as regards frauds and historical data regarding frauds in a graphical representation.

CONCLUSIONS

Fraud detection in cash transactions has been a keen area of research for the researchers for years and will be an intriguing area of research in the coming future. This happens majorly due to continuous change of patterns in frauds. In this project we propose fraud detection in cash transactions by detecting four different patterns of fraudulent transactions using best suiting algorithms and by addressing the related problems identified by past researchers in fraud detection in cash transaction. By addressing real time fraud detection in cash transaction by using predictive analytics. This part of our system can allow the fraud investigation team to make their decision to move to the next step as soon as a suspicious transaction is detected. Optimal algorithms that address four main types of frauds were selected through literature, experimenting and parameter tuning as shown in the methodology. We also assess sampling methods that effectively address the skewed distribution of data. We

can conclude that there is a major impact of using re-sampling techniques for obtaining a comparatively higher performance from the classifier. Further the models indicated 75%, 85%, 75% and 90% accuracy rates respectively. As the developed machine learning models present an average level of accuracy, we hope to focus on improving the prediction levels to acquire a better prediction. Also the future extensions aim to focus on location based frauds.

REFERENCES

1. V. Bhusari and S. Patil, "Study of hidden markov model in credit card fraudulent detection", 2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave), pp. 1-4, 2016.

2. https://www.projectpro.io/article/credit-card-fraud-detection-project-with-source-code-in-python/568

3. https://nevonprojects.com/credit-card-fraud-detection-project/

4. M. Zareapoor, P. Shamsolmoali et al., "Application of credit card fraud detection: Based on bagging ensemble classifier", Procedia computer science, vol. 48, no. 2015, pp. 679-685, 2015.

5. https://www.geeksforgeeks.org/ml-credit-card-fraud-detection/

6. https://www.insightbig.com/post/credit-card-fraud-detection-with-machine-learning-in-python

7. P. K. Chan, W. Fan, A. L. Prodromidis and S. J. Stolfo, "Distributed data mining in credit card fraud detection", IEEE Intelligent Systems and Their Applications, vol. 14, no. 6, pp. 67-74, 1999.

8. https://data-flair.training/blogs/credit-card-fraud-detection-python-machine-learning/

9. https://github.com/raviprakash11/CreditCardFraudDetectionSystem