CREDIT CARD FRAUD DETECTION USING LOGISTIC REGRESSION

K. Gowthami¹, K V L E Praneetha², G. Vinitha³, Ch. Reddamma Kumari⁴
P. Sandhya Krishna⁵

¹²³⁴ IV B.Tech, Department of Information Technology, Vignan’s Nirula Institute of Technology & Science for Women, Peda Palakaluru, Guntur-522009, Andhra Pradesh, India.
⁵Asst.Professor, Department of Information Technology, Vignan’s Nirula Institute of Technology & Science for Women, Peda Palakaluru, Guntur-522009, Andhra Pradesh, India.

sandhyarncaviit@gmail.com

ABSTRACT

Use of online transactions in day to day life has been increasing since last decade due to advancements in technology and network connectivity. Due to ease, simplicity and user friendliness of the online transaction system, new users are constantly joining the vast population benefiting from such system. Credit card fraud resulting from misuse of the system is defined as theft or misuse of one’s credit card information which is used for personal gains without the permission of the card holder. To detect such frauds, it is important to check the usage patterns of a user over the past transactions. Credit card fraud refers to the physical loss of credit card or loss of sensitive credit card information. Many machine-learning algorithms can be used for detection.

Keywords— Credit card fraud, Machine learning algorithms, logistic regression.

1. INTRODUCTION

It is important that credit card companies are able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase [1]. Credit card fraud are often outlined because the bootleg use of any system or, criminal activity through the utilization of physical card or card info while not the information of the cardholder[2][3]. The MasterCard is also physical or virtual in an exceedingly physical-card, the cardholder presents his or her card physically to a business person for creating a payment. To hold out deceitful transactions during this quite purchase, a wrongdoer needs to steal the MasterCard. Machine learning algorithms are employed to analyse all the authorized transactions and report the suspicious ones[4][5]. These reports are investigated by professionals who contact the cardholders to confirm if the transaction was genuine or fraudulent. The investigators provide a feedback to the automated system which is used to train and update the algorithm to eventually improve the fraud-detection performance over time[6][7]. Fraud detection methods are continuously developed to defend criminals in adapting to their fraudulent strategies. These frauds are classified as:

• Credit Card Frauds: Online and Offline
• Card Theft
• Account Bankruptcy
• Device Intrusion
• Application Fraud
• Counterfeit Card
• Telecommunication Fraud

Some of the currently used approaches to detection of such fraud are:

• Artificial Neural Network
• Fuzzy Logic
• Genetic Algorithm
• Logistic Regression
• Decision tree
• Support Vector Machines
• Bayesian Networks
• Hidden Markov Model
• K-Nearest Neighbour

2. LITERATURE REVIEW

Prepare Data for Logistic Regression

The assumptions made by logistic regression about the distribution and relationships in your data are much the same as the assumptions made in linear regression[8][9]. Much study has gone into defining these assumptions and precise probabilistic and statistical language is used. My advice is to use these as guidelines or rules of thumb and experiment with different data.
preparation schemes[10][11]. Ultimately in predictive modelling machine learning projects you are laser focused on making accurate predictions rather than interpreting the results[12]. As such, you can break some assumptions as long as the model is robust and performs well.

Binary Output Variable: This might be obvious as we have already mentioned it, but logistic regression is intended for binary (two-class) classification problems[13]. It will predict the probability of an instance belonging to the default class, which can be snapped into a 0 or 1 classification[14][15].

Remove Noise: Logistic regression assumes no error in the output variable (y), consider removing outliers and possibly misclassified instances from your training data[16][17]. Gaussian distribution: Logistic regression is a linear algorithm (with a non-linear transform on output[18]). It does assume a linear relationship between the input variables with the output[19][20]. Data transforms of your input variables that better expose this linear relationship can result in a more accurate model[19][43]. For example, you can use log, root, Box-Cox and other univariate transforms to better expose this relationship[20][44].

Remove Correlated Inputs: Like linear regression, the model can over fit if you have multiple highly-correlated inputs[45][46]. Consider calculating the pair wise correlations between all inputs and removing highly correlated inputs[21][22]. Fail to Converge: It is possible for the expected likelihood estimation process that learns the coefficients to fail to converge[23]. This can happen if there are many highly correlated inputs in your data or the data is very sparse (e.g. lots of zeros in your input data)[24][25].

Further Reading
There is a lot of material available on logistic regression. It is a favorite in may disciplines such as life sciences and economics[46][47].

Logistic Function
Logistic regression is named for the function used at the core of the method, the logistic function[26]. The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment[27][28]. It’s an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits[48].

\[ \frac{1}{1 + e^{-value}} \]

Where e is the base of the natural logarithms (Euler’s number or the EXP() function in your spreadsheet) and value is the actual numerical value that you want to transform[29]. Below is a plot of the numbers between -10 and 10 transformed into the range 0 and 1 using the logistic function[30][31].

![Logistic function representation](image)

3. PROPOSED SYSTEM
In this proposed system we are doing the analysis of techniques of fraud detection which is user friendly and secure[32]. This system analyzes the feasibility of credit card fraud detection and proposes this detection procedures and its empirical process.

It contains only numerical input variables which are the result of a PCA transformation[33][34]. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data[35][36]. Features V1, V2... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset[37]. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-sensitive learning[38]. Feature 'Class' is the response variable and it takes negative value in case of fraud and positive value otherwise[39][40].

Implementation code
creditcard.drop(columns = 'time', inplace = True)
# Normalize the 'amount' column
scaler = StandardScaler()
creditcard['amount'] = scaler.fit_transform(creditcard['amount'].values.reshape((-1, 1)))
# creditcard.drop(columns = 'amount', inplace = True)
X = creditcard.iloc[:, :-1]
y = creditcard.iloc[:, -1]
Xtrain, Xtest, ytrain, ytest = train_test_split(X, y, test_size = .33, stratify = y, random_state = seed)
logitreg_parameters = {'C': np.power(10.0, np.arange(-3, 3))}
logitreg = LogisticRegression(verbose = 3, warm_start = True)
logitreg_grid = GridSearchCV(logitreg, param_grid = logitreg_parameters, scoring = 'roc_auc', n_jobs = 70)
logitreg_grid.fit(Xtrain, ytrain)   
logitreg_grid.best_params_ ;

4. RESULTS
The code prints out the number of fraud cases and no fraud cases detected and compares it with the actual values. This is used to calculate the accuracy of the different algorithms. The complete dataset is also used for testing and the results are printed. These results along with the classification report for algorithm is given in the output as follows, the positive values in the result gives no fraud occurred and the negative values in the result gives the fraud occurred.

Result screens python

Fig 2  open the file in python GUI

Fig 3 enter value1 and value2

Fig 4 Negative value results the fraud occurred

Fig 5 Positive values in results no fraud occurred
5. CONCLUSION

Fraud detection is a complicated issue, requires more planning and practice to implement with machine learning algorithms. This makes sure that the customer’s cash is secure and no longer without difficulty tampered with. Usage of credit cards is increased in each field of day to day life. Credit card Frauds also be increased. To enhance security of the transaction systems an effective way efficient credit card detection system is established.

REFERENCES


[36]. VellalacheruPavani and I. Ramesh Babu, “A Novel Method to Optimize the


