Data Analytics based Credit card Fraud Detection system using Logistic Regression

Abstract:

In the present scenario with the great improvements in technology, credit cards are used for online purchasing and cause sudden outbreak in credit card fraud. Fraud detection is concerned with not only capturing the fraudulent events, but also capturing of such activities as quickly as possible. In the existing credit card fraud detection system, fraudulent transaction will be detected after transaction is done. Thus, fraud is spreading all over the world, resulting in huge financial losses. Our project presents an analysis to compare the performance of "Logistic Regression" method in credit card fraud detection with a data set. At the same time, this project trying to ensure that genuine transactions are not rejected.

Keywords: Logistic Regression, credit card, fraud Detection, Fraud.

1. INTRODUCTION

Credit cards are used for purchasing goods and services with the help of virtual card, which is used for online transaction. Generally, fraud can be defined as criminal deception or illegal activity in financial or personal gain to damage another individual without necessarily leading to direct legal consequences. Credit card fraud, a wide ranging term for theft and fraud committed or any similar payment mechanism as a fraudulent resource of funds in a transaction.

With the increased use of credit cards, fraudsters are also finding more opportunities to fraudulent activities which effects bank as well as card holders to large financial losses. Fraud detection based on analysing existing purchase data of cardholder is a promising way for reducing the credit card frauds.

In this study, a credit card fraud detection system using Logistic Regression method is developed. In this system, each account is monitored separately using descriptors, and the transactions are identified and flagged as fraud or normal.

2. TYPES OF FRAUD

Various types of frauds like credit card frauds, Bankruptcy fraud, Theft fraud, Application fraud, Behavioural fraud.

1. Credit card Fraud: It involves two types of frauds.
    1. Online fraud
    2. Offline fraud

    Online is committed through phone, Internet, shopping or absence of card holder.

    Offline is committed by using a stolen physical card at call center or any other place.
2. Bankruptcy Fraud:

    It is one of the most complicated types of fraud to predict. The bank will send its users/customers an order to pay. The users will be recognized as being in state of personal bankruptcy and not able to recover their unwanted loans.

3. Theft Fraud:

It refers using a card that is not yours. The owner give some feedback and contact the bank, the bank will take measures to check the thief as early as possible.

4. Application Fraud:

When someone applies for a credit card with false information is termed as application fraud.

5. Behavioural Fraud:

It occurs when sales are made on 'cardholder' present basis and details of legitimate cards have been obtain fraudulent basis.

3. RELATED WORK

Credit card fraud detection has drawn a lot of research interest and involves eaves dropping on the behaviour of users for detecting or avoid undesirable behaviour of customers.

In previous research, Ghosh and Reilly[1] have proposed credit card fraud detection with a neural network. .Recently, Syed et al.[2]have used parallel granular neural networks (PGNNs) for improving the speed of data mining and knowledge discovery process in credit card fraud detection. Stolfo et al. [3] suggest a credit card fraud detection system (FDS) using meta learning techniques to learn models of fraudulent credit card transactions. Meta learning is a general strategy that provides a means for combining and integrating a number of separately built classifiers or models. Aleskerov et al. [4] present CARDWATCH, a database mining system used for credit card fraud detection. The system, based on a neural learning module, provides an interface to a variety of commercial databases. Kim and Kim [5] have identified skewed distribution of data and mix of legitimate and fraudulent transactions as the two main reasons for the complexity of credit card fraud detection .Based on this observation, they use fraud density of real transaction data as a confidence value and generate the weighted fraud score . Fan et al. [6] suggest the application of distributed data mining in credit card fraud detection. Brause et al. [7] have developed an approach that involves advanced data mining techniques and neural network algorithms to obtain high fraud coverage. Chiu and Tsai [8] have proposed Web services and data mining techniques to establish a collaborative scheme for fraud detection in the banking industry. Phua et al. [9] have done an extensive survey of existing data-mining-based FDSs and published a comprehensive report. Prodromidis and Stolfo [10] use an agent-based approach with distributed learning for detecting frauds in credit card transactions. Phua et al. [11] suggest the use of meta classifier in fraud detection problems. They consider naive Bayesian, C4.5, and Back Propagation neural networks as the base classifiers. Vatsa et al. [12] have recently proposed a game-theoretic approach to credit card fraud detection. They model the interaction between an attacker and an FDS as a multi stage game between two players, each trying to maximize his payoff. Ourston et al. [13] have proposed the application of HMM in detecting multistage network attacks. Hoang et al. [14] present a new method to process sequences of system calls for anomaly detection using HMM. Lane [15] has used HMM to model human behaviour.

This project proposes Logistic Regression [16] algorithm, to solve the credit card fraud problem . It gives more accurate results when compared to previous models.

4. PROPOSED SYSTEM

LOGISTIC REGRESSION:

Logistic regression is a statistical method for analysing dataset in which there are one or more independent variables that determine an outcome.

It is a regression model where the dependent variable is categorical. The binary logistic model is used to estimate the probability of a binary response based on one or more predictor.

It is useful for situations in which we want to be able to predict the presence or absence of outcome based on value of a set of predictor variables. Logistic regression coefficients can be used to estimate odds ratios for each of independent variables in model and it is applicable to broader range of situations for estimating the odds with probability.

$$\ln\left(\frac{p}{1-p}\right)=b_0+b_1 \bullet x$$

log (p/1-p) is the link function. Logarithmic transformation on the outcome variable allows us to model a non-linear association in a linear way.

CONFUSION MATRIX

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing.

For example confusion matrix for a binary classifier (though it can easily be extended to the case of more than two classes):

n=165	**Predicted: NO**	**Predicted: YES**
Actual: NO	50	10
Actual: YES	5	100

- true positives (TP): These are cases in which we predicted yes (they have the disease), and they do have the disease.
- true negatives (TN): We predicted no, and they don't have the disease.
- false positives (FP): We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")
- false negatives (FN): We predicted no, but they actually do have the disease. (Also known as a "Type II error.")

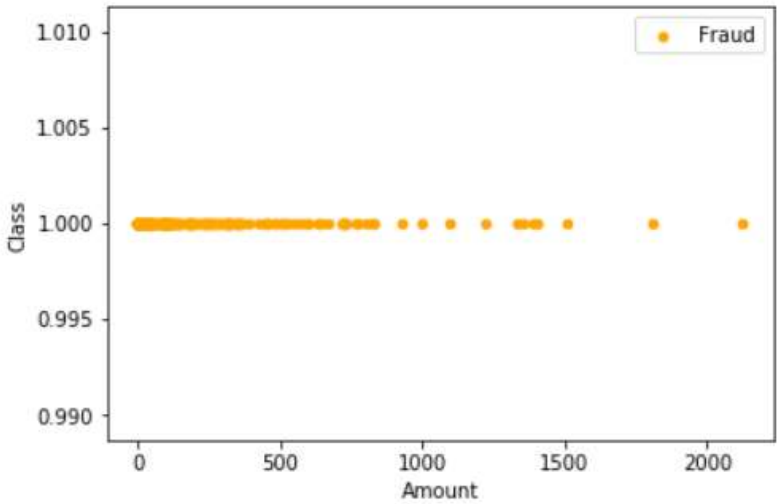
I've added these terms to the confusion matrix, and also added the row and column totals:

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

This is a list of rates that are often computed from a confusion matrix for a binary classifier:
Accuracy: Overall, how often is the classifier correct?
(TP+TN)/total = (100+50)/165 = 0.91

Misclassification Rate: Overall, how often is it wrong?
(FP +FN)/total = (10+5)/165 = 0.09
True Positive Rate: When it's actually yes, how often does it predict yes?
TP/actual yes = 100/105 = 0.95
False Positive Rate: When it's actually no, how often does it predict yes?
FP/actual no = 10/60 = 0.17
Specificity: When it's actually no, how often does it predict no?
TN/actual no = 50/60 = 0.83
Precision: When it predicts yes, how often is it correct?
TP/predicted yes = 100/110 = 0.91
Prevalence: How often does the yes condition actually occur in our sample?
actual yes/total = 105/165 = 0.64

In our project, we can calculate accuracy using binary confusion matrix as follows:



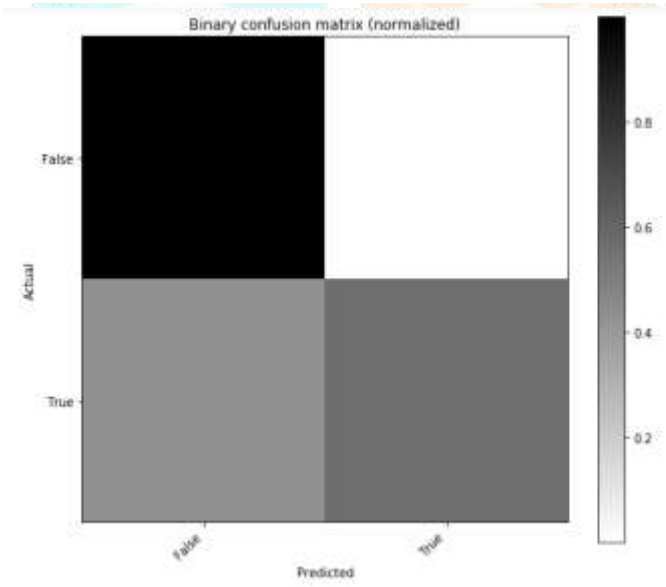
In the above graph, it clearly shows the fraud transactions are marked with orange colour based on the amount parameter.

```
logistic = linear_model.LogisticRegression(C=1e5)
logistic.fit(X_train, y_train)
print("Score: ", logistic.score(X_test, y_test))
```

```
Score:  0.998986788118
```

```
from pandas_ml import ConfusionMatrix
confusion_matrix = ConfusionMatrix(y_right, y_predicted)
print("Confusion matrix:\n%s" % confusion_matrix)
confusion_matrix.plot(normalized=True)
plt.show()
confusion_matrix.print_stats()
```

```
Confusion matrix:
Predicted  False  True  __all__
Actual
False      99474    38    99512
True          74    97      171
__all__    99548   135    99683
```



```
population: 99683
P: 171
N: 99512
PositiveTest: 135
NegativeTest: 99548
TP: 97
TN: 99474
FP: 38
FN: 74
TPR: 0.567251461988
TNR: 0.999618136506
PPV: 0.718518518519
NPV: 0.999256640013
FPR: 0.00038186349385
FDR: 0.281481481481
FNR: 0.432748538012
ACC: 0.998876438309
F1_score: 0.633986928105
MCC: 0.637875313777
informedness: 0.566869598494
markedness: 0.717775158531
prevalence: 0.00171543793826
LRP: 1485.48230225
LRN: 0.432913852008
DOR: 3431.35775249
FOR: 0.000743359987142
```

5. CONCLUSION

To improve security of the transaction systems in effective way, building an accurate and efficient credit card fraud detection system is the key task for institutions. The different steps in credit card transaction processing are represented as the underlying process of LR. It is also explained whether a transaction is fraud or not. The system is also scalable for handling large volumes of transactions.

6. FUTURE ENHANCEMENT

As a future work, instead of making performance comparisons just over the prediction accuracy, We will alert the user before a transaction is going to be taken place.

7. ACKNOWLEDGEMENT

Firstly, We would like to express our sincere gratitude towards our project guide Dr. Indraneel Sreeram for valuable guidance, suggestions and encouragement.

8. REFERENCES

[1]. Ghosh, S., and Reilly, D.L., 1994. Credit Card Fraud Detection with a Neural-Network, 27th Hawaii International l Conference on Information Systems, vol. 3 (2003), pp. 621- 630.

[2]. Syeda, M., Zhang, Y. Q., and Pan, Y., 2002 Parallel Granular Networks for Fast Credit Card Fraud Detection, Proceedings of IEEE International Conference on Fuzzy Systems, pp. 572577 (2002).

[3]. Stolfo, S. J., Fan, D. W., Lee, W., Prodromidis, A., and Chan, P. K., 2000. Cost-Based Modeling for Fraud and Intrusion Detection: Results from the JAM Project, Proceedings of DARPA Information Survivability Conference and Exposition, vol. 2 (2000), pp. 130-144.

[4]. Aleskerov, E., Freisleben, B., and Rao, B., 1997. CARDWATCH: A Neural Network Based Database Mining System for Credit Card Fraud Detection, Proceedings of IEEE/IAFE: Computational Intelligence for Financial Eng. (1997), pp. 220-226.

[5]. M.J. Kim and T.S. Kim, "A Neural Classifier with Fraud Density Map for Effective Credit Card Fraud Detection," Proc. Int'l Conf. Intelligent Data Eng. and Automated Learning, pp. 378-383, 2002.

[6]. W. Fan, A.L. Prodromidis, and S.J. Stolfo, "Distributed Data Mining in Credit Card Fraud Detection," IEEE Intelligent Systems, vol. 14, no. 6, pp. 67-74, 1999.

[7]. R. Brause, T. Langsdorf, and M. Hepp, "Neural Data Mining for Credit Card Fraud Detection," Proc. IEEE Int'l Conf. Tools with Artificial Intelligence, pp. 103-106, 1999. [8]. C. Chiu and C. Tsai, "A Web Services-Based Collaborative Scheme for Credit Card Fraud Detection," Proc. IEEE Int'l Conf. e-Technology, e-Commerce and e Service, pp. 177181, 2004.

[9]. C. Phua, V. Lee, K. Smith, and R. Gayler, "A Comprehensive Survey of Data Mining-Based Fraud Detection Research, http://www.bsys.monash.edu.au/people/cphua/. Mar. 2007.

[10]. S. Stolfo and A.L. Prodromidis, "Agent-Based Distributed Learning Applied to Fraud Detection," Technical Report CUCS-014-99, Columbia Univ., 1999.

[11]. C. Phua, D. Alahakoon, and V. Lee, "Minority Report in Fraud Detection: Classification of Skewed Data," ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 50-59, 2004.

[12]. V. Vatsa, S. Sural, and A.K. Majumdar, "A Game-theoretic Approach to Credit Card Fraud Detection," Proc. First Int'l Conf. Information Systems Security, pp. 263-276, 2005

[13]. D. Ourston, S. Matzner, W. Stump, and B. Hopkins, "Applications of Hidden Markov Models to Detecting Multi Stage Network Attacks," Proc. 36th Ann. Hawaii Int'l Conf. System Sciences, vol. 9, pp. 334-344, 2003.

[14]. X.D. Hoang, J. Hu, and P. Bertok, "A Multi-Layer Model for Anomaly Intrusion Detection Using Program Sequences of System Calls," Proc. 11th IEEE Int'l Conf. Networks, pp. 531-536, 2003.

[15]. T. Lane, "Hidden Markov Models for Human/Computer Interface Modelling," Proc. Int'l Joint Conf. Artificial Intelligence, Workshop Learning about Users, pp. 35-44, 199

[16]. Y. Sahin, E. Duman, Detecting Credit Card Fraud by ANN and Logistic Regression, ©2011 IEEE.