How to Cite:
Ajithkumar, A., & Geetha, S. (2022). Classification of e-commerce financial transaction logs using machine learning approach. *International Journal of Health Sciences*, *6*(S1), 4500–4506. https://doi.org/10.53730/ijhs.v6nS1.5835

Classification of e-commerce financial transaction logs using machine learning approach

A Ajithkumar
Department of Computer Applications, MepcoSchlenk Engineering College, Sivakasi
Email: ajithkumar6001@gmail.com

S Geetha
Assistant Professor (Sr. Grade), Department of Computer Applications, MepcoSchlenk Engineering College, Sivakasi
Email: sgeetha@mepcoeng.ac.in

***Abstract*---**E-Commerce becomes inevitable in the current world. Especially in this pandemic period, almost activities have been carried out through digital mode to serve the customers with more safety. The financial transactions are much secured in the e-commerce payment. Still, many intruders are making these transactions into fail and hacking the customers' information when the financial transactions are carried out. Along with, sometimes the transactions may get failed due to the network issues. Hence, the E-Commerce organizations are maintaining the transaction logs to check and take necessary action over the failed transactions. Out of huge transaction logs, identifying the suspicious failed transactions in manual method is not encouraged. Lot of technologies have come to support the detection of suspicious failed transactions such as Artificial Neural Network, Machine Learning, Deep Learning and other statistical methods. As the above methods are proved as more suitable for detecting and classifying the failed transaction logs, still the accuracy of classification has not been achieved much, which motivated to propose the machine learning based classification of E-Commerce Financial Transaction Logs. In this paper, the optimized classification of financial transaction logs are done using the Logistic Regression and Support Vector Machine approach and the result shows the higher accuracy of classification in Support motivated to propose the machine learning based classification of E-Commerce Financial Transaction Logs. In this paper, the optimized classification of financial transaction logs are done using the Logistic Regression and Support Vector Machine approach and the result shows the higher

International Journal of Health Sciences ISSN 2550-6978 E-ISSN 2550-696X © 2022.
***Corresponding author**: Geetha, S.; Email: sgeetha@mepcoeng.ac.in*
Manuscript submitted: 27 Feb 2022, Manuscript revised: 18 March 2022, Accepted for publication: 09 April 2022

accuracy of classification in Support Vector Machine approach when compared with the logistic regression.

***Keywords*---**Financial Transaction logs, Log Classification, E-Commerce, Logistic Regression, Support Vector Machine.

Introduction

In this digital world everything is getting to digitalized form under that now a day's amount transaction, payments and other bank transactions can be done with the help of internet using some gateways and third party applications or sites these are commonly known as online transaction. But we are still facing some issues during the online transactions. During the online transaction more and more different types of errors and failures may occur. It is difficult to find out all the errors and rectify it. Sometimes particular error or failure occurs frequently to all the users.

Large amount of log data are generating each day, containing various sensitive and invaluable information. Data analysts use this information to improve the user's experience including the system's security and confidentiality. As the amount of data doubles every day, it is becoming more challenging for data analysts to track errors on different servers and also training the resources. The system analysts commonly use the grep command on the error log messages to find out the server failures or application failures and in several cases if the error message is irrelevant, it is difficult to figure out what to search for. The keyword as "FAILURE" and "ERROR" are not helpful for data analysts as these error messages may occurs during system maintenance or actual application server failure and the terms are interchangeable.

Error messages do not occur in a similar pattern and sometimes go through different stages and accumulate extra information that makes analysing log data more challenging. The current IT operation system at the City of Calgary comprises of different servers such as the applications server and the monitoring server which generate thousands of error logs, and collectively get stored in databases.

Here the error or failure data are stored as log file in the xml format, and the data in the xml file is extracted using python and visualized using the method. It will help to find out the error which occurs frequently with these major errors can be rectified and the online transactions become more reliable.

Materials and Methods

Related Works

An Unsupervised anomaly detection framework works with a HDFS log files and successfully detect anomalies with 83%[1]. There is different algorithm for anomaly detection [2] to identify the anomaly in the logs and found that Local Outlier Factor is performed better when compared to other metrics. The purpose

extracting data from the logs by optimizing window sizes without any lose of information [3]. The result is not just visualizing the log data and also provides insight into the logs through topics from the error messages.

E-commerce has developed rapidly and is currently in a golden age of digital economy. E-commerce has been utilized byeach individual or company to the days to sell or get merchandise and services in the form of electronic payment [4]. It is difficult for an e-commerce system to deal with fraud, because a component is often shared by multiple software, and many threats and attacks [5]. Total-order-based model is used to represent the logical relationship of transaction record attributes [6]. The heuristic supervised algorithm has developed to identify the difference between the real behaviours and the behaviours determined by models [7][9,10]. The Novel method has been used for detecting abnormal transactions via an integration model of data and control flows and it successfully detect some comprehensive analysis [8].

Dataset

The dataset is collected as financial transaction logs from accounts management system for e-commerce. The dataset contains nearly 5000 logs about the successful as well as unsuccessful transactions. It contains various attributes such as customer id, invoice id, amount applied, payment mode, payment description, transaction date, account id. The dataset is in xml format.

Pre-processing

The collected data set contains many missing data which has to be processed to make it into the complete dataset. The dataset with missing values may not give proper results in the specified neural network model when it is used for training. The missing values are imputed with interpolation methods and normalized to reduce the discrepancy among the various data.

Implementation

The pre-processed dataset is divided into training set and testing set. The 70% of records have been used for training and remaining 30% records have been used for testing in the both Logistic Regression as well as Support Vector Machine. The training data set will be used to training the classification model to learn the features and it will be generate the model fit to classify the transaction logs in to specific class.

Logistic regression

The alternate for the linear regression is Logistic when the values are given with categorical nature. In Logistic Regression, the sigmoid function takes the real input and bring the output as probability value 0 to 1. Sigmoid function can be represented as below. Let y is the dependent variable; p(y) is the probability of the dependent variable y; s is the Sigmoid Function; nlog is the natural logarithm;

$$s[p(x)] = nlog\left(\frac{p(x)}{1-p(x)}\right) = \alpha + \beta x,\ where\ odds = \left(\frac{p(x)}{1-p(x)}\right) \quad (1)$$

Where β is the variable used to specify the rate of increase or decrease of the S-shaped curve of p(x).

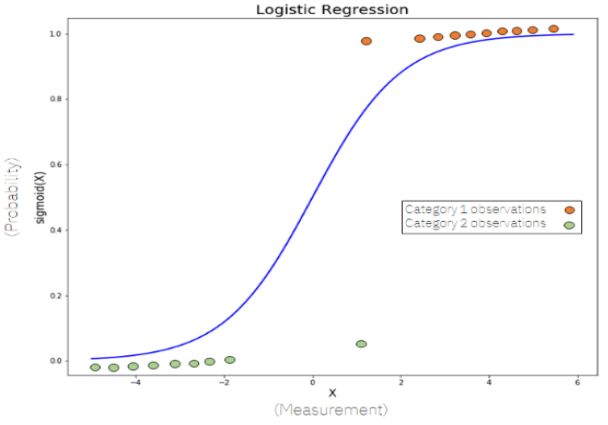

Figure 1. Logistic Regression

Support Vector Machine

Support Vector Machine (SVM) is the machine learning algorithm mainly used for binary classification. In SVM, the hyperplane is constructed where it took the form of mapping input space to support non-linear classification problem and find the best hyperplane for the give data.

$$q.x + p = 0,\ where\ q\ is\ weight\ vector,\ p\ is\ bias \quad (2)$$

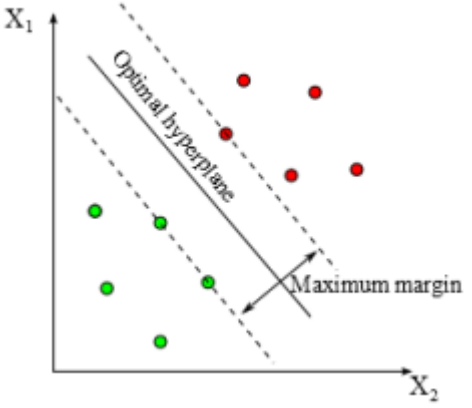

Figure 2. Best Hyperplane using SVM

Evaluation measures

In classification problems, many algorithms are available to classify the data into two classes which is called as binary classification and to classify the data into multi class which is called as multiclass classification. The classification algorithms can be evaluated in various ways to identify the developed model is fit and provide the efficient and accurate classification of the given dataset. In this proposed work, the following evaluation measures have been used to evaluate the classification model.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Accuracy Score to divide number of correct predictions by total number of predictions and get the percentage of samples were predicted correctly

$$\text{True Positive Rate} = \frac{\text{TP}}{\text{TP+FN}} \quad (4)$$

True Positive Rate compares the number of correct predictions that the transactions is completed with the total number of boxes with alive cats

$$True\ Negative\ Rate = \frac{TN}{TN+FP} \quad (5)$$

True Negative Rate compares the number of boxes with failed transactions and the number of correct predictions that the cat is dead:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (6)$$

It shows the ability of our model to differ the “Successful transactions” class from all the others, but, unfortunately, it does not give an idea of whether we have found all the Transactions are failed or successful not.

Another evaluation measure is the Recall which is used to identify the proportion of failure transaction logs specified by:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (7)$$

The F1 score is another evaluation measure which considers the average between precision. It is defined by:

$$\text{F1 Score} = 2\left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}}\right) \quad (8)$$

Results and Discussions

The given data set has been visualized in the Fig. 3 to understand the various types of transactions occurring and logged into the transaction files. And the success rates all the transactions are visualized in the Fig. 4. The Table 1 shows the evaluation measures applied on both logistic regression and support vector machine and obtained values. The support vector machine provides better accuracy than the logistic regression

Fig. 3: Transactions through all the methods

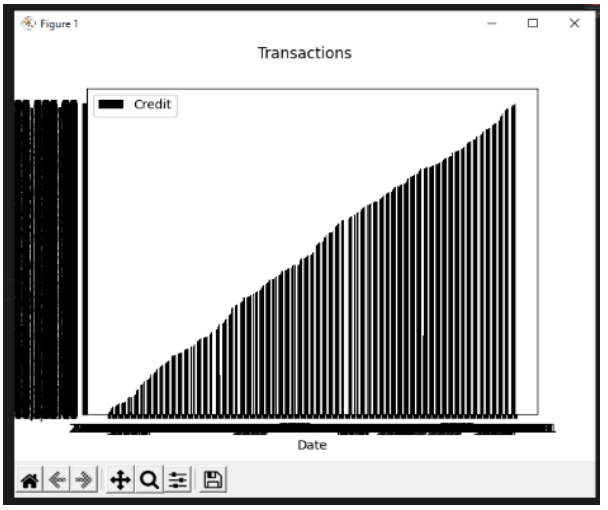

Fig 4: Success rate of Transaction done

Table 1
Evaluation Results

Approach	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Logistic Regression	97.3	100	92.3	96
Support VectorMachine	97.6	100	94.1	97

Conclusion

In our approach, we have proposed visualizing the Transaction events through anomaly detection. In this paper, we dealt with transaction error logs classification and we have reduced the efforts of analysing transaction errors, and the visualization of error logs shows the promising results. In addition to the results, our approach provides the sequence of events which can be used for data flow. We are exploring the different set of log sets from open source in order to generalize our approach. In our future work, we will visualize and classify the error logs with more accuracy and it may help to find out the better solution to solve the problems.

Acknowledgments

I obliged to give my appreciation to a number of people without whom I could not have completed this thesis successfully. I would like to place on record my deep sense of gratitude and thanks to my internal guide Mrs.S.Geetha, Department of Computer Applications, Mepco Schlenk Engineering College, Sivakasi, whose esteemed support and immense guidance encouraged me to complete the project successfully. I would like to thank our HoD Dr. P.Radha, Department of Computer Applications, Mepco Schlenk Engineering College, Sivakasi for their valuable support and encouragement to take up and complete this work.

References

[1] Zeufack, Vannel, "An Unsupervised Anomaly Detection Framework for Detecting Anomalies in Real Time through Network System's Log Files Analysis", 2020.

[2] Sri Sai Manoj Kommineni, AkhilaDindi, "Automating Log Analysis", Blekinge Institute of Technology, Faculty of Computing, Department of Computer Science 2021.

[3] Suman, R, "An Approach to Server Log Analysis for Abnormal Behaviour Detection", University of Calgary, Calgary, AB 2021.

[4] S. Fatonah, A. Yulandari, and F. W. Wibowo, "A review of e-payment system in e-commerce," *Journal of Physics: Conference Series. IOP Publishing*, vol. 1140, no. 1, article 012033, 2018.

[5] A. Abdallah, M. A. Maarof, and A. Zainal, "*Fraud detection system: a survey*," *Journal of Network and Computer Applications*, vol. 68, pp. 90–113, 2016.

[6] L. Zheng, G. Liu, C. Yan, and C. Jiang, "Transaction fraud detection based on total order relation and behavior diversity," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 3, pp. 796–806, 2018.

[7] A. Rozinat and W. M. P. Van der Aalst, "Conformance checking of processes based on monitoring real behavior," *Information Systems*, vol. 33, no. 1, pp. 64–95, 2008.

[8] Yadi Wang, Wangyang Yu, Peng Teng, Guanjun Liu, and Dongming Xiang ,"A Detection Method for Abnormal Transactions in E-Commerce Based on Extended Data Flow Conformance Checking", Volume 2022 , Article ID 4434714,04 Jan 2022.

[9] Raja Sekar, Jayaram & Selvakumar, N. & Radha, P. & Johnsonselva, J.V., " Supervised and Unsupervised learning for characterizing the industrial material defects", International Journal of Business Intelligence and Data Mining, 1(1):1, 2022.

[10] Maheswari, K. & Rajesh, S., "A novel QIM-DCT based fusion approach for classification of remote sensing images via PSO and SVM models", Soft Computing, 24, 2020.