IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Linked Feature Set With Enhanced Logistic Regression For Online Payment Fraud Classification

Kothapalli Mandakini[1]
Research Scholar, Department of CSE
Osmania University, Hyderabad

K.Shyamala[2]
Professor, Department of CSE
Osmania University, Hyderabad.

Abstract

Online Payment Methods are becoming popular both in-store and online. Online payment fraud occurs frequently due to the open nature of the internet, which allows criminals to use certain techniques to commit fraud, such as eavesdropping, phishing, infiltration, denial-of-service, database theft, and man-in-the-middle assault. Making a secured system for client authentication and fraud prevention is difficult since there is always a way around it. This means that fraud detection systems play a crucial role in avoiding financial misdeeds.Recognizing dishonest financial dealings is made significantly easier with the use of machine learning and statistical methods. Due to the sensitive nature of the data, it is challenging to make inferences and develop more accurate models. This research proposes a Linked Feature Set with Enhanced Logistic Regression (LFS-ELR) Model for accurate classification of online payment frauds. The proposed model classification accuracy is high when contrasted with the existing models.

Keywords:. Classification Enhanced Logistic Regression. Feature Subset, Fraud Detection, Machine Learning, Online Payment system

1. Introduction

Fraud occurs in all areas of commerce, including online shopping, healthcare, banking, and finance. Almost a trillion dollars in annual revenue is gained by fraud. While fraud poses serious risks to businesses, sophisticated tools like rules engines and machine learning can assist discover instances of it [8]. Online payments fraud has skyrocketed along with the expansion of business that accepts its payments. In this paper, the use of machine learning for spotting fraudulent activity is proposed. In order to detect and prevent online payment fraud, the suggested system use enhanced logistic regression to construct a classifier [9]. The goal of this research has been to develop a model that uses machine learning to identify this kind of online payment fraud [14].Classification is an important part of machine learning, and logistic regression is one method that can be used for that purpose. A logistic function is used to represent the independent variable. Because of its binary nature, the dependent variable can take on just two distinct values. This study makes use of regression analysis, a relatively new technique in the field of data analytics for massive online payment datasets [15]. It is clear from the literature that regression methods can only be used with small datasets containing a few hundred entries at most. So, it is not an easy process to apply regression to large datasets [17]. Regression analyses are favored since they are predicated solely on the interdependence of the variables of interest [18]. Fraud detection of online payment systemis dynamic test bed for researchers to improve accuracy of the models.

2. Related Work

Online payment fraud is the most concerning irregularities that can occur during a transaction. Data mining techniques are one of the many methods that researchers have been exploring as a possible solution to these problems. It can be difficult for researchers to make sense of the credit card information that has been obtained. Two datasets were employed in this investigation by Kalid et al. [2] credit card frauds (CCF) and credit card default payments (CCDP). Successful anomaly detection is achieved by the MCS's sequential decision combination technique. Hashemi et al. [3] optimizing the hyperparameters, the author took into account real-world factors like imbalanced data and employed a method called Bayesian optimization. To improve the performance of the LightGBM technique, as well as CatBoost and XGBoost to take into consideration the voting mechanism. Lastly, the author proposed a weight-tuning hyperparameter and applied deep learning to fine-tune the other hyperparameters in order to further increase performance.

Using the IBM Safer Payments and IBM Quantum Computers with the Qiskit software stack, Grossi et al. [4] provided a quantum support vector machine (QSVM) algorithm for a classification task. The author conducted a comprehensive comparison machine-learning and an ensemble model combining classical and quantum algorithms is investigated and providing a hybrid classical-quantum strategy. The author discovered that the outcomes are quite sensitive to the feature selections and methods employed.Alarfaj et al. [6] developed a method for detecting frauds,focused on a high-class imbalanced data, The strategies include the Extreme Learning Method, Decision Tree, Random Forest, Support Vector Machine, Logistic Regression, and XG Boost. The detection accuracy was further enhanced by the inclusion of layers.Due to the extreme imbalanced of data sets, machine learning approaches are ineffective in combating fraud. An improved version of the Support Vector Machine (SVM) using quantum annealing solvers has been used to implement the detection framework presented in this paper by Wang et al. [7]. QML application's detection performance was compared to that of twelve machine learning implementations on two datasets.The results validate the efficacy of conventional machine learning techniques for non-time series data, while highlighting the promise of QML applications for time series based, highly unbalanced data.

Using real-world imbalanced datasets, Ileberi et al. [8] implemented a machine learning (ML) based approach for credit card fraud detection. Synthetic Minority over-sampling Technique(SMOTE) was used to resample the dataset. And Support Vector Machine, Logistic Regression, Random Forest, Extreme Gradient Boosting, Decision Tree, and Extra Tree were among the ML techniques used to assess the framework (ET). To further improve their accuracy at classifying data, these ML algorithms were combined with the Adaptive Boosting (AdaBoost) method. Measures such as the Matthews Correlation Coefficient (MCC) and Area Under the Curve (AUC) were used to assess the models' performance (AUC). In addition, the suggested methodology was applied to a highly skewed synthetic credit card fraud dataset to further verify the findings of this study.

The process of feature selection has been viewed as a useful tool for addressing unbalanced classification issues. Finding a small feature subset that yields a high classification accuracy can be posed as a multiobjective optimization problem (MOP). Traditional MOP focuses on determining an optimal solution while disregarding the variety of possible solutions. In this paper, Han et al. [9] adopted a multimodal MOP (MMOP) perspective on feature selection, with the objectives being to locate a strong Pareto front in objective space and to discover as many analogous Pareto optimal solutions . In order to solve this problem, a brand new competition-driven technique has been developed to aid existing multimodal MMEAs in discovering more comparable feature subsets as well as a desired Pareto front.

3. Proposed Model

Credit card transactions in India grew from 390 million in 2012 to 652 million in 2021.The dynamic and ever-evolving nature of the financial services industry and the high stakes involved, fraudsters are able to take advantage of several chances. The rise in online payment fraud in recent years has had a significant impact on the economy. First and foremost, there is feature selection, which involves picking out the most important variables from a dataset.That's why it's important that the datasets used to train machine learning models for spotting fraud exclude personally identifying information from their properties. The proposed model framework is shown in Figure 2.

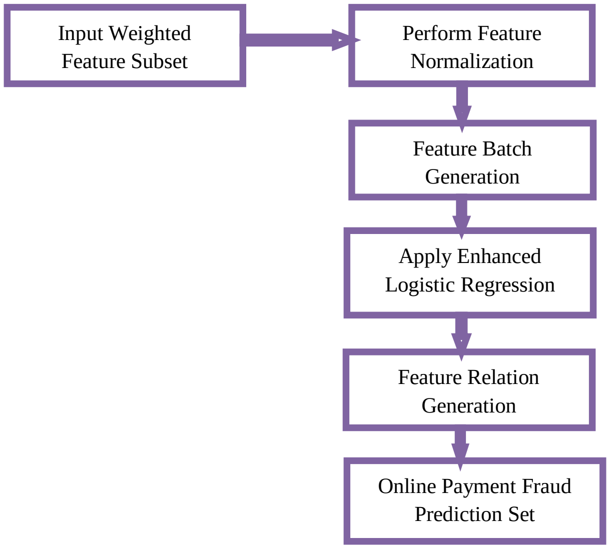
Input Weighted Feature Subset
Perform Feature Normalization
Feature Batch Generation
Apply Enhanced Logistic Regression
Feature Relation Generation
Online Payment Fraud Prediction Set

Figure 2: Proposed Model Framework

Predicting a logic of the dependent variable from a set of explanatory variables is the goal of logistic regression. This research proposes a Linked Feature Set with Enhanced Logistic Regression (LFS-ELR) Model for accurate classification of online payment frauds. The enhanced logistic regression model performs the regression analysis on dependent variables at multiple levels for reduced feature set generation. The proposed LFS-ELR algorithm is clearly discussed in this section.

Input: Weighted Feature Vector {WFVset}

Output: Online Payment Fraud Data Set {OPFDset}

Step-1: The weighted features are considered from the dataset and analyzed for processing individual features for fraud detection. The feature loading and processing is performed as

$$FeatSet(WFVset[M]) = \sum_{f=1}^{M} \frac{\max\Big(FeatSet\big(Corr(f,f+1)\big)\Big)-\min\Big(FeatSet\big(Corr(f,f+1)\big)\Big)+\max(Wfeat(f))}{\text{len}(\text{FeatSet}[\text{M}])} \quad (1)$$

Here FeatSet is the weighted feature set considered for processing. Max feature correlation set is considered and remaining are excluded that has less weight.

Step-2: The weighted feature set is input and the features are analyzed for performing the feature normalization. The objective of the normalization process is to adjust the values of the features.

$$FeatNorm(FeatSet[M]) = \prod_{f=1}^{M} getmean(FeatSet(f)) - \sum_{f=1} \frac{std(FeatSet(f,f+1)))}{\sqrt{\text{median}(\max(\text{Featset}(\text{f}))}} \quad (2)$$

Here getmean() is used to find the mean values of the features to perform normalization by balancing the feature values, std is used to calculate the standard deviation among the considered feature set.

Step-3: The identifying information that batch generators provide is referred to as batch feature generation. The feature batch generation is performed as

$$FeatBatch[\text{M}] = \left(\sum_{f=1} \sum_{corr=0} \frac{\text{getmin}(\text{FeatNorm}(\text{f}))}{\lambda} + \lim_{\text{f}\to WFVset} \left(\gamma + \frac{\max(\text{FeatNorm}(\text{f+1}))}{\text{median}(\max(\text{Featset}(\text{f}))}\right)^2)\right) \quad (3)$$

λ is the batch size considered based on the normalized feature set. γ is the model used for batch vector processing of features that are normalized as maximum.

Step-4: Enhanced Logistic Regression modeling assumes that the log-odds of an event are linear combinations of one or even more independent variables to estimate its likelihood.The ELR model considers the independent features and maps these features with the batch feature set and the multi-level independent feature set is generated as

$$Lreg = \log\left\{\frac{\max(FeatBatch(f))}{\lambda}\right\} + maxset(corr(f, f+1)) \quad (4)$$

$$\mathrm{ELreg} = \lambda_{\max}(Lreg) + \sqrt{\sum_{f=1}^{M} \min(\mathrm{FeatBatch(f+1)}) + \sum_{f=1}^{M} \frac{\max corr(f+1, M-f)}{mean(Lreg(f))}}$$

Step-5: The relational coefficient between two features is 1 if they depend on one another in a linear fashion. A relational coefficient of 0 indicates that there is no relationship between the features. The process of feature relation generation is performed as

$$Frel[M] = \sum_{f=1}^{M} \frac{mindiff(f, f+1)}{std(ELreg(f))}$$

$$\mathrm{UFrel[M]} = \frac{\sum_{\mathrm{f=1}}^{\mathrm{R}} \delta(\mathrm{Frel(f,f+1)}) - \min(\mathrm{diff}(\mathrm{Frel(f,f+1)}))}{\sum_{\mathrm{f=1}}^{\mathrm{M}} \gamma * \mathrm{size(FeatBatch)}}$$

(5)

Step-6: The ELR model and the final online payment fraud set is generated based on the feature relations trained and the final prediction set is generated as

$$PSet(UFrel[M]) = \sum_{f=1}^{M} diff(UFrel(\max(Frel(f+1, f)) + \sum_{f=1}^{M} \frac{\lambda(UFrel(f)) + \sum_{i=1}^{N} \mathrm{Frel(f+1)}}{\mathrm{size(UFrel)}}$$

(6)

4. Results

Criminals are more likely to resort to online payment fraud in an attempt to overcome the security measures of payment providers as the popularity of such transactions has skyrocketed. With the ultimate goal of preventing frauds on a online payment system and developing countermeasures against attacks, there is consequently a great deal of pressure to research potential security concerns that may be exploited. Spotting potentially fraudulent financial dealings in their infancy is a crucial part of this research. With the rise of online payment services, there has been a corresponding increase in the demand for automated detection methods that can spot and stop fraudulent transactions in real time. This research proposes a Linked Feature Set with Enhanced Logistic Regression (LFS-ELR) Model for accurate classification of online payment frauds. The proposed model is compared with the traditional Fine-Grained Co-Occurrences for Behavior-Based Fraud Detection in Online Payment Services (FGCO-BFD-OPS) and the results show that the proposed model performance is high. The fraud prediction set is calculated using the formulas specified below.

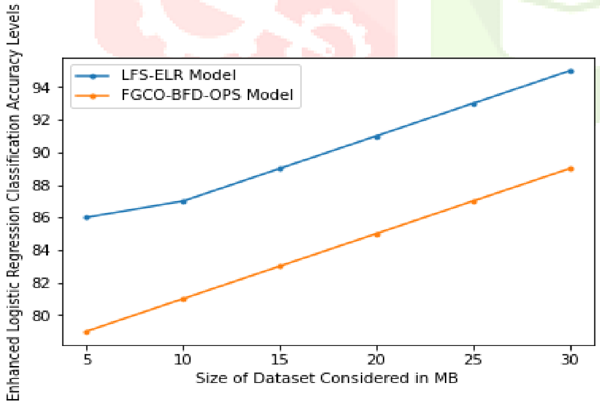

Fig 8: Enhanced Logistic Regression Classification Accuracy Levels

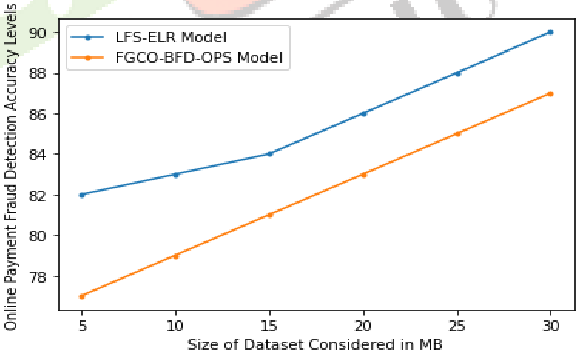

Fig 9: Online Payment Fraud Detection Accuracy Levels

Table 3.1: Comparison of Results

S.No	DS-Size(mb)	Model comparision	PP-Accuracy (%)	FE-Time levels (ms)	FE-Acc levels(%)	Corr.cal Acc levels(%)	FW-Allo time levels(ms)	FW-Allo Acc levels(%)
1	5	ML-AWFV	90.12	5	86	88	9	82
	5	FGCBFD	84.92	11	79	78	15	78
2	10	ML-AWFV	90.91	7	86.92	89.12	11	83.61
	10	FGCBFD	87	12.5	81	80.91	17	80.12
3	15	ML-AWFV	92.30	9	88.56	90	12	86
	15	FGCBFD	87.51	14.8	81.65	82.12	18	81.21
4	20	ML-AWFV	94.82	10	91	92.15	14	87.82
	20	FGCBFD	88.02	16	82.51	83.54	19	83.81
5	25	ML-AWFV	96.15	11	93	94.35	14	89.25
	25	FGCBFD	91.13	18	85.21	85.31	21	84.01
6	30	ML-AWFV	99.02	14	95.62	97.50	15	92
	30	FGCBFD	92.21	21	86.10	86.21	24	86.10

Note: DS-Data set, PP-Pre-Processsing,FE-Feature Extraction,FW.Allo-Feature Weight Allocation

5. Conclusion

Online Payments fraud is one of the most common forms of online payment fraud, although it may happen to anyone using any payment method. Due to the sheer volume of daily transactions, manual fraud detection checks are just not feasible. Thus, such systems require swiftness and precision in their development. Machine learning helps for detecting online payment fraud, but only if the right features are employed. This research proposes a Linked Feature Set with Enhanced Logistic Regression(LFS-ELR) Model for accurate classification of online payment frauds. The proposed model not only functions in these settings, but also delivers higher precision, which ultimately results in less waste and lower costs. In future, integrated classifiers can be designed to reduce the feature set and to increase the precision rate and optimization can also be embedded into this model for better performance accuracy is 99.44%.

References

1. C. Wang and H. Zhu, "Representing Fine-Grained Co-Occurrences for Behavior-Based Fraud Detection in Online Payment Services," in IEEE Transactions on Dependable and Secure Computing, vol. 19, no. 1, pp. 301-315, 1 Jan.-Feb. 2022, doi: 10.1109/TDSC.2020.2991872.
2. S. N. Kalid, K. -H. Ng, G. -K. Tong and K. -C. Khor, "A Multiple Classifiers System for Anomaly Detection in Credit Card Data With Unbalanced and Overlapped Classes," in IEEE Access, vol. 8, pp. 28210-28221, 2020, doi: 10.1109/ACCESS.2020.2972009.
3. S. K. Hashemi, S. L. Mirtaheri and S. Greco, "Fraud Detection in Banking Data by Machine Learning Techniques," in IEEE Access, vol. 11, pp. 3034-3043, 2023, doi: 10.1109/ACCESS.2022.3232287.
4. M. Grossi et al., "Mixed Quantum–Classical Method for Fraud Detection With Quantum Feature Selection," in IEEE Transactions on Quantum Engineering, vol. 3, pp. 1-12, 2022, Art no. 3102812, doi: 10.1109/TQE.2022.3213474.
5. E. Esenogho, I. D. Mienye, T. G. Swart, K. Aruleba and G. Obaido, "A Neural Network Ensemble With Feature Engineering for Improved Credit Card Fraud Detection," in IEEE Access, vol. 10, pp. 16400-16407, 2022, doi: 10.1109/ACCESS.2022.3148298.

6. F. K. Alarfaj, I. Malik, H. U. Khan, N. Almusallam, M. Ramzan and M. Ahmed, "Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms," in IEEE Access, vol. 10, pp. 39700-39715, 2022, doi: 10.1109/ACCESS.2022.3166891.
7. H. Wang, W. Wang, Y. Liu and B. Alidaee, "Integrating Machine Learning Algorithms With Quantum Annealing Solvers for Online Fraud Detection," in IEEE Access, vol. 10, pp. 75908-75917, 2022, doi: 10.1109/ACCESS.2022.3190897.
8. E. Ileberi, Y. Sun and Z. Wang, "Performance Evaluation of Machine Learning Methods for Credit Card Fraud Detection Using SMOTE and AdaBoost," in IEEE Access, vol. 9, pp. 165286-165294, 2021, doi: 10.1109/ACCESS.2021.3134330.
9. S. Han, K. Zhu, M. Zhou and X. Cai, "Competition-Driven Multimodal Multiobjective Optimization and Its Application to Feature Selection for Credit Card Fraud Detection," in IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 52, no. 12, pp. 7845-7857, Dec. 2022, doi: 10.1109/TSMC.2022.3171549.
10. Y. -Y. Hsin, T. -S. Dai, Y. -W. Ti, M. -C. Huang, T. -H. Chiang and L. -C. Liu, "Feature Engineering and Resampling Strategies for Fund Transfer Fraud With Limited Transaction Data and a Time-Inhomogeneous Modi Operandi," in IEEE Access, vol. 10, pp. 86101-86116, 2022, doi: 10.1109/ACCESS.2022.3199425.
11. B. Baesens, S. Höppner and T. Verdonck, "Data engineering for fraud detection", Decis. Support Syst., vol. 150, Nov. 2021.
12. X. Zhang, Y. Han, W. Xu and Q. Wang, "HOBA: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture", Inf. Sci., vol. 557, pp. 302-316, May 2021.
13. Y. Xie, G. Liu, R. Cao, Z. Li, C. Yan and C. Jiang, "A feature extraction method for credit card fraud detection", Proc. 2nd Int. Conf. Intell. Auton. Syst. (ICoIAS), pp. 70-75, 2019.
14. E. U. Savona and M. Riccardi, "Assessing the risk of money laundering: Research challenges and implications for practitioners", Eur. J. Criminal Policy Res., vol. 25, no. 1, pp. 1-4, Mar. 2019.
15. D. Vassallo, V. Vella and J. Ellul, "Application of gradient boosting algorithms for anti-money laundering in cryptocurrencies", Social Netw. Comput. Sci., vol. 2, no. 3, May 2021.
16. H. Zhu, G. Liu, M. Zhou, Y. Xie, A. Abusorrah and Q. Kang, "Optimizing weighted extreme learning machines for imbalanced classification and application to credit card fraud detection", Neurocomputing, vol. 407, pp. 50-62, Sep. 2020.
17. T. Zhang, K. Zhu and D. Niyato, "A generative adversarial learning-based approach for cell outage detection in self-organizing cellular networks", IEEE Wireless Commun. Lett., vol. 9, no. 2, pp. 171-174, Feb. 2020.
18. P. Zhang, S. Shu and M. Zhou, "An online fault detection model and strategies based on SVM-grid in clouds", IEEE/CAA J. Automatica Sinica, vol. 5, no. 2, pp. 445-456, Mar. 2018.