CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING

Dr. G. Sivakumar Prof./CSE[1], Ms. S. Haridharani[2], Ms. J. Jerrine Varshini[3], Ms.P.Keerthana[4]

Assistant Professor[1], Student[2 3 4], Department of Computer Science and Engineering

Erode Sengunthar Engineering College, Perundurai, Erode, Tamil Nadu, India.

ABSTRACT

Credit card fraud has become a major problem in the financial industry, necessitating the implementation of effective fraud detection systems. In this paper, we investigate the use of three machine learning methods for credit card fraud detection: Naive Bayes, Logistic Regression, oneR, and the J48 (C4.5) decision tree. The study uses historical credit card transaction data that includes both valid and fraudulent purchases. Following pre-processing and feature engineering, the data is separated into training and testing sets. On the testing set, we train each model and evaluate its performance using a variety of measures such as accuracy, precision, and recall, as well as F1-score and ROC-AUC. The findings demonstrate the efficacy of each algorithm in detecting credit card fraud. Comparisons across the models illustrate their various strengths and shortcomings, offering useful information for determining the best method. Furthermore, we investigate the prospect of increasing detection accuracy using ensemble approaches such as Voting Classifier, Bagging, or Boosting to improve the overall fraud detection system. Finally, the study intends to help financial institutions choose appropriate machine learning algorithms for credit card fraud detection, highlighting the necessity of constant monitoring and model modification in combating emerging fraud strategies efficiently.

Keywords: Credit card, machine learning, fraud detection, banking sectors.

1. INTRODUCTION

1.1 CREDIT CARD

Credit cards have transformed how we make purchases and manage our finances. A credit card is a plastic payment card provided by a financial organization, usually a bank, that allows the cardholder to borrow money to purchase goods and services. Unlike debit cards, which remove cash straight from the cardholder's bank account, credit cards provide a line of credit that must be returned at a later date, typically with interest. The cardholder may use the credit card for a variety of purchases, both in person and online, making it a convenient and generally recognized payment method across the world. Credit cards have become an essential aspect of modern consumer culture, giving people more purchasing power and the freedom to manage their finances properly. When a consumer applies for a credit card and is authorized, the financial institution sets them a credit limit, which is the most money they may borrow with the card. The cardholder may use the credit card to make purchases up to the limit. Each month, the credit card firm delivers a billing statement outlining the transactions that occurred during that time period as well as the minimum amount owing. While paying the minimal amount keeps the account in good standing, it is best to pay off the whole debt to prevent interest costs. If the cardholder fails to make the minimum payment on time, they risk incurring late penalties and harming their credit score. Credit cards sometimes come with a variety of incentives and advantages, such as cash back, travel points, or discounts, to encourage card use and loyalty.

1.2 FRAUD DETECTION

Fraud detection is a vital use of data analysis and machine learning that aims to identify and prevent fraudulent activity and transactions across several domains. As technology progresses and financial transactions move to digital platforms, the danger of fraudulent activity increases, making fraud detection more important than ever. Whether it's credit card fraud, insurance fraud, identity theft, or internet scams, companies and financial institutions use sophisticated fraud detection systems to protect their assets, consumers, and preserve faith in their services. Fraud detection systems examine massive volumes of transactional and behavioural data to find anomalies and suspect trends. Fraud may have serious consequences, including financial losses, tarnished reputations, and impaired consumer trust.

2. LITERATURE REVIEW

2.1 UNCERTAINTY- AWARE CREDIT CARD FRAUD DETECTION USING DEEP LEARNING

Maryam Habibpour [1] and colleagues propose in this work Numerous research have used deep neural networks (DNNs) to identify credit card fraud, with the goal of improving point prediction accuracy and avoiding unwanted biases through the development of various network architectures or learning models. It is critical to measure uncertainty in conjunction with point estimate because it lowers model unfairness and allows practitioners to build dependable systems that prevent making incorrect judgments due to uncertainty. Because fraudsters continuously change their strategies, DNNs meet observations that do not come from the same process as the training distribution. Furthermore, because to the time-consuming nature of the procedure, only few transactions are reviewed by experienced specialists in order to update DNNs. These characteristics make it necessary to clearly evaluate the uncertainty associated with DNN predictions in real-world card fraud detection scenarios.

2.2 CREDIT CARD FRAUD DETECTION IN THE ERA OF DISRUPTIVE TECHNOLOGIES: A SYSTEMATIC REVIEW

Credit card fraud is becoming a serious and growing problem as new technologies and communication channels emerge, such as contactless payment. This article gives a comprehensive examination of current research on detecting and forecasting fraudulent credit card transactions from

2015 to 2021. The 40 papers picked for consideration are analysed and classified based on the topics they cover (class imbalance problem, feature engineering, etc.) and the machine learning approach they use (conventional and deep learning modeling). Our analysis reveals that deep learning has received little research, implying that more research is needed to address the difficulties in detecting credit card fraud using cutting-edge technologies such as big data analytics, large-scale machine learning, and cloud computing. Our study is a significant resource for academic and industrial researchers in analysing financial fraud detection systems and designing trustworthy solutions by highlighting existing research challenges and future research opportunities.

2.3 MACHINE LEARNING BASED ON RESAMPLING APPROACHES AND DEEP REINFORCEMENT LEARNING FOR CREDIT CARD DETECTION SYSTEMS

Dr. Tran Khanh Dang [3], et. The issue of unbalanced datasets is a fundamental concern for constructing reliable credit card fraud (CCF) detection systems, as stated in this system. In this paper, we study and evaluate current advances in deep reinforcement learning (DRL) and machine learning (ML) algorithms for CCF detection systems, including fraud and non-fraud labels. The imbalanced CCF dataset is resampled with SMOTE and ADASYN, two resampling methods. This balanced dataset is then exposed to ML algorithms to generate CCF detection models. The imbalanced CCF dataset is then used to build detection algorithms with DRL.

2.4 ON THE BLACK-BOX CHALLENGE FOR FRAUD DETECTION USING MACHINE LEARNING (II) NONLINEAR ANALYSIS THROUGH INTERPRETABLE AUTOENCODERS

Jacobo Chaquet-Ulldemolins [4] et al. proposed this system. Artificial intelligence (AI) has lately gained popularity in the global economy due to its exceptional ability to analyse and model data in a variety of disciplines. As a result of this condition, society is rapidly becoming more automated, and these new approaches may be combined to form a beneficial tool for addressing the difficult challenge of credit fraud detection. However, severe restrictions make it impossible for financial institutions to comply with them while employing modern procedures. From a methodological approach, auto encoders have demonstrated effectiveness in finding nonlinear features in a range of problem domains. However, auto encoders are opaque and often referred to as "black boxes." In this study, we provide an interpretable and impartial CFD approach.

2.5 CREDIT CARD FRAUD DETECTION USING A NEW HYBRID MACHINE LEARNING ARCHITECTURE

Esraa Faisal Malik [5], for example. As mentioned in this article Financial crimes have progressively harmed financial institutions. Various single and hybrid machine learning algorithms have been used to detect crimes such as credit card fraud. However, due to a lack of additional research on alternative hybrid algorithms for a specific dataset, these techniques have significant limitations. This paper proposes and tests seven hybrid machine learning models for detecting fraudulent acts on a real-world dataset. Modern machine learning techniques were initially applied to detect credit card fraud, and the best single algorithm from the first phase was used to create the hybrid approaches. The hybrid models created were separated into two phases. Our results revealed that the hybrid model AdaBoost + LGBM is the best model due to its superior performance. Future study should focus on exploring different hybridization strategies and credit card domain algorithms.

3. RELATED WORK

In today's digital economy, credit cards are indispensable, and as their use has recently increased significantly, so has credit card theft. Algorithms for machine learning (ML) have been used to detect credit card fraud. However, it has proven challenging for ML classifiers to function at their best due to credit card holders' dynamic shopping habits and the issue of class imbalance. This paper presents a robust deep-learning method to address this issue, which combines a multilayer perceptron (MLP) as the meta-learner with long short-term memory (LSTM) and gated recurrent unit (GRU) neural networks as base learners in a stacking ensemble architecture. To balance the class distribution in the dataset, the hybrid synthetic minority oversampling methodology and edited nearest neighbor (SMOTE-ENN) method are used. According to the experimental findings, the suggested deep learning ensemble in combination with the SMOTE-ENN method produced sensitivity and specificity values of 1.000 and 0.997, respectively, which are better than those of other commonly employed ML classifiers and techniques in the literature.

4. METHODOLOGY

The historical credit card transaction data collection, The data pre-processing step entails gathering historical credit card transaction data and undertaking extensive data cleaning and feature engineering. Relevant information, such as transaction amount, location, time of day, and cardholder behavior, are retrieved to build a large dataset appropriate for training the models. To improve the accuracy and resilience of the fraud detection system, we use ensemble modeling approaches. This entails merging the results of multiple machine learning algorithms, such as Naive Bayes, Logistic Regression, oneR, and the J48 (C4.5) decision tree, utilizing methods like Voting Classifier or Bagging. By combining predictions from various models, we hope to decrease false positives and false negatives, resulting in more reliable and efficient fraud detection. The suggested system's fundamental function is real-time fraud detection. Once trained and refined, the ensemble model is deployed in a real-time credit card transaction processing system. As new transactions occur, the model quickly analyses their risk level and issues alerts for possible fraudulent activity. The system's real-time nature allows for quick replies, reducing possible losses and providing a safe experience for cardholders.

A. Load Data

The goal of the feature selection module is to choose the pre-processed data's most pertinent and instructive features. To determine the most important characteristics that contribute to the detection of credit card fraud, it makes use of a variety of methodologies, including statistical testing, correlation analysis, or feature importance rankings. The credit card transaction data must first be loaded into a framework for machine learning. Several sources, including financial institutions, credit card firms, and publicly accessible databases, can provide the data.

B. Data Pre-Processing

The initial cleaning and preparing of the

$$LR(z) = \frac{1}{1+e^{z}}$$

credit card transaction data is handled by the Data Pre-processing module. It carries out operations such handling missing values, identifying and treating outliers, and normalizing or scaling data. It also finds and fixes any problems with the quality of the data that can affect how accurate the future models are.

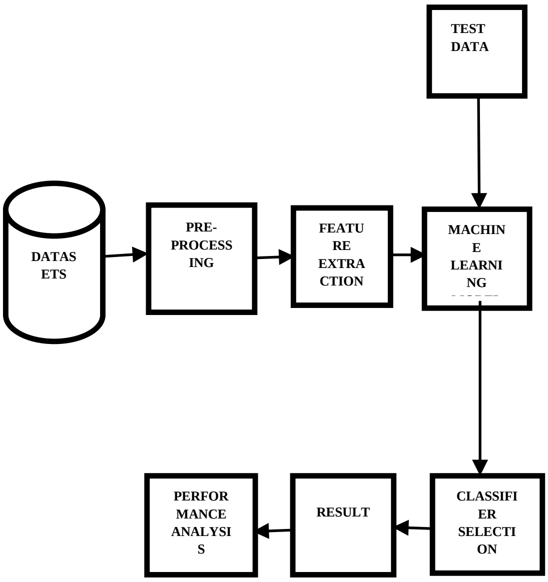

Figure 1. Block diagram

5. ALGORITHM DETAILS

A. Naïve Bayes Algorithm

Naïve Bayes algorithm is a simple probabilistic algorithm in classification technique which gets its probability value based on the calculation of frequency and value combinations from the associated collection. This algorithm assumes that all attributes are independent The classification process of Naïve Bayes demands several clues or directions to determine the class of the data to be analyzed. Therefore, Equation 1 is applied

$$P(C|F_1...F_n)=\frac{P(C).P(F_1...F_n|C)}{P(F_1...F_n)}$$

In Equation 1, variable *C* is a class, and variable *F1...Fn* represents characteristics required to do a classification. Therefore, the probability of matching data with a certain characteristic in C class (posterior) is C class probability emerged multiplied by the probability of samples characteristics in C class (likelihood), and then divided by the probability of samples characteristics globally (evidence).

Input:

Training dataset T,

F= (f1, f2, f3,.., fn) // value of the predictor variable in testing dataset.

Output:

A class of testing dataset.

Step:

1. Read the training dataset T;

2. Calculate the mean and standard deviation of the predictor variables in each class;

3. Repeat

Calculate the probability of *fi* using the gauss density equation in each class;

Until the probability of all predictor variables (f1, f2, f3,.., fn) has been calculated.

4. Calculate the likelihood for each class;

5. Get the greatest likelihood;

B. Logistic Regression

Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable (or output), y, can take only discrete values for given set of features (or inputs), X.

Logistic regression becomes a classification technique only when a decision threshold is brought into the picture. The setting of the threshold value is a very important aspect of Logistic regression and is dependent on the classification problem itself. As per feature selection used for the data set here the best threshold value for logistic Regression is 0.6

C. J48 decision tree

It is a predictive method to analyze the target value from a dataset on various given attributes. From the training data, it finds the attribute which segregate several instances. In order to achieve highest information gain, these instances are further classified. This procedure is applied over the smaller subsets in a repetitive manner until all the instances rightly placed in their class. In the given figure 1, the first level is a single header node which is a pointing node to its children. Attributes are denoted by internal nodes whereas the branches give possible values these attributes can have.

6. CONCLUSION

The credit card fraud detection system, which uses Naïve Bayes, Logistic Regression, and J48 (C4.5) algorithms, provides a comprehensive and robust strategy to prevent fraudulent financial transactions. The system's goal is to identify fraudulent transactions with high accuracy, precision, recall, and F1-score while minimizing false positives by meticulous data pre-processing, feature selection, and ensemble modeling. If a real-time fraud detection module is added, it improves the system's efficacy by immediately raising alarms for possibly fraudulent transactions, allowing for early action to avert losses. The user-friendly interface allows customers to configure the system's parameters, allowing for specialized fraud detection depending on unique needs. With continual monitoring and updates, the system can react to developing fraud tendencies, ensuring a high degree of security in an ever-changing financial environment.

7. REFERENCES

[1] M. Habibpour, H. Gharoun, M. Mehdipour, A. Tajally, H. Asgharnezhad, A. Shamsi, A. Khosravi, M. Shafie-Khah, S. Nahavandi, and J. P. S. Catalao, "Uncertainty-aware credit card fraud detection using deep learning 2021; arXiv:2107.13508.

[2] A. Cherif, A. Badhib, H. Ammar, S. Alshehri, M. Kalkatawi, and A. Imine. "Credit card fraud detection in the era of disruptive technologies: A systematic review." J. King Saud Univ. Computer and Information Science, vol. 35, no. 1, pp. 145-174, Jan. 2023, doi:10.1016/j.jksuci.2022.11.008.

[3] T. K. Dang, T. C. Tran, L. M. Tuan, and M. V. Tiep. "Machine learning based on resampling approaches and deep reinforcement learning for credit card fraud detection systems." Appl. Sci., vol. 11, no. 21, p. 10004, Oct. 2021; doi: 10.3390/app112110004.

[4] Chaquet-Ulldemolins et al., "On the black-box problem for fraud detection using machine learning (I): Linear models and informative feature selection," Applied Sciences, vol. 12, no. 7, p. 3328, March 2022, doi: 10.3390/app12073328.

[5] E. F. Malik, K. W. Khaw, B. Belaton, W. P. Wong, and X. Chew. "Credit card fraud detection using a new hybrid machine learning architecture." Mathematics, vol. 10, no. 9, p. 1480, April 2022; doi: 10.3390/math10091480.

[6] I. Benchaji, S. Douzi, B. El Ouahidi, and J. Jaafari, "Enhanced credit card fraud detection using attention mechanism and LSTM deep model," J. Big Data, vol. 8, no. 1, p. 151, December 2021; doi: 10.1186/s40537-021-00541-8.

[7] 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." IEEE Access, vol. 10, pp. 16400–16407, 2022; doi: 10.1109/ACCESS.2022.3148298.

[8] E. Btoush, X. Zhou, R. Gururaian, K. Chan, and X. Tao, "A survey on credit card fraud detection approaches in banking sector for cyber security," in Proc. 8th International Conf. Behav. Social Comput. (BESC), Oct. 2021, pp. 1-7, doi: 10.1109/BESC53957.2021.9635559.

[9] Y. Xie, G. Liu, C. Yan, C. Jiang, M. Zhou, and M. Li, "Learning transactional behavioral representations for credit card fraud detection IEEE Trans. Neural Networks and Learning Systems, early access, October 5, 2022, doi: 10.1109/TNNLS.2022.3208967.

[10] J. Yang & J. Guan "A heart disease prediction model based on feature optimization and the smote-Xgboost algorithm," Information, vol. 13, no. 10, p. 475, Oct. 2022, doi: 10.3390/info13100475.