Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 2 (2024)

Credit Card Fraud Detection Using Hybrid Machine Learning Algorithms

Prajna Parimita Nayak [1], Dr. Balaji Madhavan [2], Umadevi G [3]

[1] *Student, Department of Computer Science and Engineering, Agni College of Technology, Chennai, Tamil Nadu, India*

[2] *Head of Department, Department of Computer Science and Engineering, Agni College of Technology, Chennai, Tamil Nadu, India*

[3] *Assistant Professor, Department of Computer Science and Engineering, Agni College of Technology, Chennai, Tamil Nadu, India*

Abstract:- Evolution and civilization are two faces of the same coin. Human civilization has evolved from the stone age to the iron age and currently in 2024 we are living in the digital age. Now we have easy access to the internet and electronic devices. Improved cost of living, abundant availability of internet and gadgets has facilitated the use of cashless transactions. Nowadays we can pay using net banking, mobile banking, mobile apps, tap and swipe of cards. It increases the possibility of fraud in financial transactions. Fraud is a phenomenon when we do certain things without the consent of authorized stakeholders. Now a days physical presence of a credit card is not necessary, we can transact if we have details of the credit card. So, this increases the possibility of credit card fraud. To stop credit card fraud, we should have a strong fraud prevention and detection mechanism. Credit card fraud is a loss for both the issuer as well as the user. In this article, we are going to discuss a fraud detection mechanism using Machine learning algorithms. Fraud detection is a binary classification problem where we have to classify the transactions as fraud or non-fraud. In this article, we are using a hybrid ensemble voting classifier for fraud detection. Logistic Regression, Decision Tree and K Nearest Neighbor algorithm are being used to set up the hybrid model. Ensemble voting classifiers consist of several homogeneous weak learners and the prediction capability of the Ensemble hybrid model is better than individual models. Sometimes the hybrid machine learning model out performs legacy systems. When the precision of the model is high it predicts the fraud correctly.so in our work we are building a hybrid model whose precision is 0.97, where the individual models have the precision Logistic regression (LR) 0.94, Decision Tree (DT) 0.85 and K nearest Neighbor (KNN) 0.93.

Keywords*:* hybrid ensemble voting classifier, Logistic regression, decision tree, KNN, hybrid model, credit card fraud detection.

1. Introduction

Fraud in the finance sector weakens the economy and brings fear in the minds of the public about the safety of their wealth. Despite strong prevention mechanisms, fraudsters find a way to do fraud. So there has to be a strong fraud detection mechanism to detect frauds. Legacy systems used rule-based mechanisms to detect frauds. Rule-based systems are programmed in C/C++ and less passionate about detecting new and unseen transaction types, also the Rule-based system finds it difficult to manage a high volume of transactions. Legacy systems are static and monolithic in nature and do not welcome change in the system easily. Machine learning algorithms are data-driven algorithms, they do not rely on predefined rules rather they learn from data and predict for the future. Machine learning models can predict for new and unseen data. It is capable of handling high speed and high volume of data. Machine learning models are trained with data, then the performance of the model is evaluated. Once the model is ready to predict it is deployed in production. There are four categories of machine learning algorithms supervised, unsupervised, semi-supervised and reinforcement algorithms. In our work, we are using supervised machine learning techniques such as Logistic Regression (LR), Decision tree (DT), and K-nearest Neighbor (KNN) algorithm. In the case of supervised machine learning techniques, data is labeled, the models are employed for classification and regression problems. Logistic regression is a binary classification algorithm which

is probabilistic in nature that means it gives the likelihood of an event belonging to a class. Logistic regression is the most efficient binary classification algorithm. Binary classification means we have two classes either 0 or 1 here, if it is fraud it is 1 if non-fraud then 0. Decision tree algorithm is a supervised machine learning technique which is used for both regression and classification problems. Decision tree is represented in a tree-like structure where nodes represent tests on features, branch is the value of the attribute and leaf node is the outcome. KNN is a supervised non-parametric algorithm which is used for regression and classification problems. KNN algorithm considers K nearest neighbors of the new data and assigns the new data to the class of majority of neighbors. Nowadays hybrid machine learning model is a trend in industry. Hybrid models perform better than classical machine learning models. Hybrid models combine several individual models of the same type and build an ensemble model. The hybrid model always outperforms its individual counterpart. Logistic regression, decision tree, and KNN models are capable of detecting credit card fraud when they are implemented alone. So, in this work, we are combining LR, DT, and KNN to build a Hybrid model to detect credit card fraud.

2. Literature Survey

In their work "Credit Card Fraud Detection Using Logistic Regression" M Devika[2] et al proposed a model for credit card fraud detection. They used gradient boosting which is very expensive to implement so we are proposing a model with a voting classifier.

Anish Mahajan, Vivek Singh Baghel[1] et al described credit card fraud detection in their work "Credit Card Fraud Detection using Logistic Regression with Imbalanced Dataset". In their work, they are proposing the model with a 94% accuracy score but in our work, we are assuring a 97% accuracy score of the prediction model.

Ananya Singhai [4]et al proposed a model in their work "A Novel Methodology for Credit Card Fraud Detection using KNN Dependent Machine Learning Methodology". In financial fraud detection techniques, high precision scale is very crucial so the power of the KNN algorithm can be leveraged with other learners like LR and DT. In our proposed model, we are

Improving KNN model capacity by building a hybrid model comprising LR, DT, and KNN.

Asifuddin Nasiruddin Ahmed [5]et al proposed a method to detect fraud in their work "Detection of Credit Card Fraudulent Transactions Utilizing Machine Learning Algorithms". The shortcoming of this work is that it uses the SVM algorithm. Generally, SVM does not work well on large datasets. It is more sensitive to outliers and the choice of the right kernel function also impacts model behavior. SVM separates data on the hyperplane and does not give any probabilistic output. SVM works well in high dimension space but in credit card transactions, the dimension of the dataset is fixed and limited, so implementing SVM is not healthy. By time transaction dataset also increasing so SVM is not a good choice. So, we are proposing our work which includes Logistic Regression, KNN and DT algorithms.

Asifuddin Nasiruddin Ahmed [6]et al have done a survey work on different machine learning models in their work "A Survey on Detection of Fraudulent Credit Card Transactions Using Machine Learning Algorithms" showed that ensemble models XGBoost and Random Forest perform well. But both XGBoost and Random Forest are computationally expensive, so in our work we are using an ensemble voting classifier with Logistic Regression, K-Nearest Neighbor and Decision Tree which is very robust and cost-effective.

In their work "Evaluation of Supervised Machine Learning Algorithms for Credit Card Fraud Detection: A Comparison", Negar Nasiri[3] et al discussed Logistic regression and decision tree model individually and proposed a hybrid concept to enhance the efficiency of fraud prediction. In our work, we are improving the predictability of this model by adding an extra classifier KNN.

3. Opportunities

Machine learning vs Rule-based system

Rule based systems are preprogrammed and do not provide flexibility for implementing new changes. Also, rule-based systems are not capable of detecting new unseen transaction type. Machine learning models are data-driven

systems capable of prediction based on historical data. Machine learning models use statistical algorithms for fraud detection. Efficiency of model increases with time by learning more with new data.

Classic machine learning models vs Hybrid machine learning models

In our work, we are developing a hybrid model that is made up of individual classic models. The hybrid model combines the power of individual learners to make an efficient hybrid model. The hybrid model always outperforms the classic model. Hybrid model prediction is more accurate than the classic one.

Early fraud detection

Helps to detect fraud at an early stage so proper measures can be taken by businesses and users to prevent the fraud. Credit card users are increasing hence increasing the number of transactions. Machine learning models become more efficient by learning more so the model capacity increases with increased transactions, we do not need to worry about handling ever increasing datasets.

Increases brand value and customer trust

A robust fraud detection mechanism increases customer trust and helps in minimizing fraud and save society from the illegitimate use of wealth for financing illegal activity.

4. Dataset

The dataset contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions. It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. 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. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependent cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

Figure 1. Credit card transaction data set

5. Methodologies

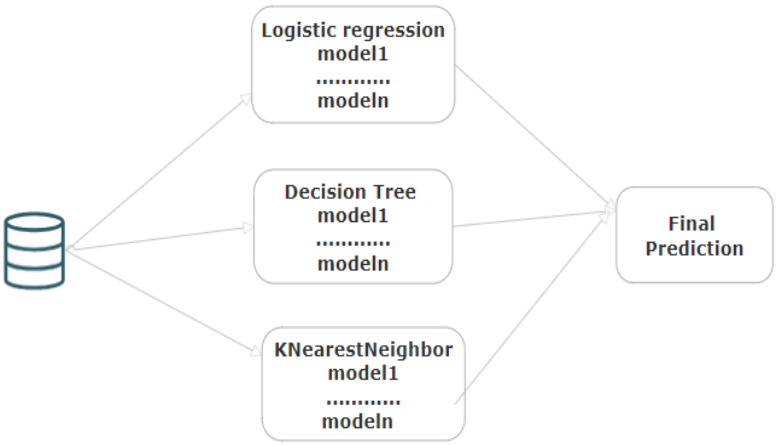

Figure 2. Hybrid Machine Learning model architecture

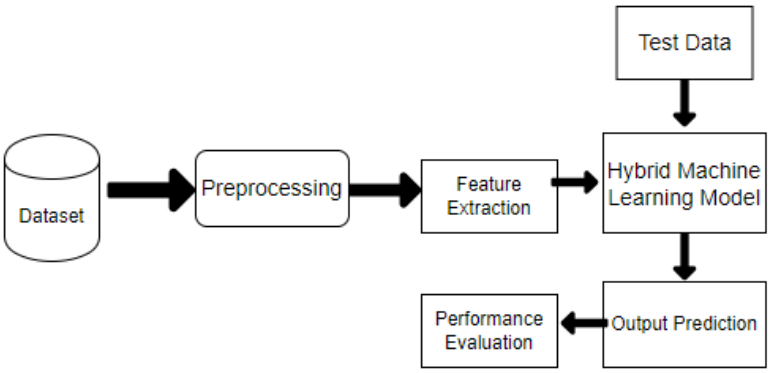

Figure 3. Machine learning Model Architecture

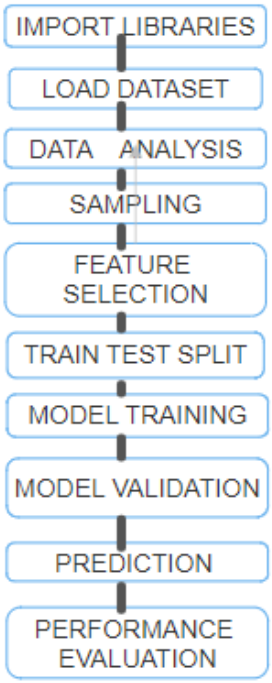

Figure 4. Control flow diagram

In this method, we are using an ensemble hybrid voting classifier which is made up of Logistic regression, Decision tree and KNN. The Hybrid algorithm architecture, control flow and model architecture is given in above figures.

Step1: Import the libraries.

Step2: Load the dataset.

Step3: Perform data analysis, separate data into two sets.

Step4: Perform under sampling to get a balanced sample so the model does not get biased.

Step5: Select feature and target from the sample.

Step6: Split data into train and test data.

Step7: Perform feature scaling.

Step8: Initialize the Machine learning model.

Step9: Train the model with training data.

Step10: Once the model is trained in above step, do prediction with test data feature set.

Step11: Once prediction is done in above step, we have to evaluate the model performance.

Step12: For hybrid ensemble model create submodels, keep the submodels in a list.

Step13: Initialize the hybrid ensemble voting classifier (soft) with submodels.

Step14: Train the hybrid model.

Step15: Do prediction for hybrid model with test data.

Step16: Evaluate performance. Hybrid model precision is 0.97 compared to LR 0. 94 , DT 0. 85 and KNN 0.93.

6. Conclusion

In conclusion, there has been much research in machine learning for fraud detection and it is growing more day by day. We tried our best to bring a robust, optimized, and efficient detection model that can detect fraud with high accuracy. A financial fraud detection system prioritizes high precision – minimizing false positives (wrongly declined transactions) to avoid inconveniencing customers. So, in our work, we are using a hybrid ensemble voting classifier that combines individual models and creates an ensemble model which is more efficient than its submodels. As the data set is imbalanced so under sampling technique has been used to get a balanced sample. We can not use all models at places. Thus, we can combine the models that are allowed and can enhance its capability using the hybrid concept. In our work we found LR having precision 0.94, DT 0.85 and KNN 0.93 but the voting classifier we implemented comes with a precision of 0.97. Other performance metrics i.e. precision, recall, confusion matrix, F1 score also calculated. We are happy with our achievement and hope to continue our research in fraud detection using Machine Learning.

7. Acknowledgement

I appreciate the support for this study from Dr.Balaji Madhvan, Head of the CSE Department, and Mrs. Umadevi G, Assistant Professor, CSE Department, Agni College of Technology, Chennai.

Reference

[1] Mahajan, A., Baghel, V. S., & Jayaraman, R. (2023, March 15-17). Credit card fraud detection using logistic regression with imbalanced dataset.

[2] Devika, M., Kishan, S. R., Manohar, L. S., & Vijaya, N. (2022, December 6-17). Credit card fraud detection using logistic regression. https://doi.org/10.1109/ICATIECE56365.2022.10046976. Retrieved from https://ieeexplore.ieee.org/document/10046976/authors#authors

[3] Mirhashemi, Q. S., Nasiri, N., & Keyvanpour, M. R. (2023, May 3-4). Evaluation of supervised machine learning algorithms for credit card fraud detection: A comparison. https://doi.org/10.1109/ICWR57742.2023.10139098. Retrieved from https://ieeexplore.ieee.org/document/10139098/authors#authors

[4] Singhai, A., Aanjankumar, S., & Poonkuntran, S. (2023, June 8). A novel methodology for credit card fraud detection using KNN dependent machine learning methodology. https://doi.org/10.1109/ICAAIC56838.2023.10141427. Retrieved from https://ieeexplore.ieee.org/document/10141427

[5] Ahmed, A. N., & Saini, R. (2023, March 3-5). Detection of credit card fraudulent transactions utilizing machine learning algorithms. https://doi.org/10.1109/INOCON57975.2023.10101137. Retrieved from https://ieeexplore.ieee.org/document/10101137

[6] Ahmed, A. N., & Saini, R. (2023, January 19-20). A survey on detection of fraudulent credit card transactions using machine learning algorithms. https://doi.org/10.1109/ICCT56969.2023.10076122. Retrieved from https://ieeexplore.ieee.org/document/10076122/keywords