Available online @ https://jjem.jnnce.ac.in
https: www.doi.org/10.37314/JJEM.SP0278
Indexed in International Scientific Indiexing (ISI)
Impact factor: 1.395 for 2021-22
Published on: 08 December 2024

Online Payment Fraud Detection using Machine Learning with XGBoost Classifier

Harish B. G[1], Sukruth K[2], Sachin Mulagund[3]

UBDT College of Engineering, Davangere 577 004
affiliated to Visvesvaraya Technological University, Belagavi
[1]harishbg@ubdt.org, [2]sukruthkashyap.sk@gmail.com, [3]sachinraghunathm@gmail.com

Abstract

Each year, online transaction fraud costs people and financial organizations billions of dollars, making it a serious criminal offense. Highlights the essential role played by financial institutions in identifying and mitigating fraudulent activities; Online transaction fraud can be prevented in a more proactive way using machine learning algorithms with higher precision. It is a piece of cake for someone to commit fraud regarding online transactions. Or in this case, the rise of e-commerce and other online sites have brought a plethora number of options for paying online which has also raised the danger level when it comes to getting frauded. You can easily detect the fraud in online transactions and evaluate it using machine learning methods with an increase in fraudulent activities which are reaching high rates. The focus of this project is the approach to regulate fraud detection using supervised machine learning models by analyzing former transactional data. Transactions are categorized into different groups based on transaction type. Subsequently, individual classifiers are trained and models are evaluated for accuracy. The classifier achieving the highest rating can then be selected as one of the top methods for fraud prediction. Utilizing the Kaggle Synthetic Financial Datasets for Fraud Detection dataset curated by Edgar Lopez-Rojas, we have employed a XGBoost classifier Machine Learning model for detecting fraudulent transactions. An in-depth comparison of these algorithms is conducted to determine the most effective solution.

Keywords: *Payment, Fraud, XGBoost classifier, Transactions.*

1. Introduction

The paper represents the use of XGBoost classifier for Online payment fraud detection. The current work was done on Logistic regression, Here we are using the XGBoost to make it more predictive.

Training Data: The portion of the data that our model is trained on. This is the real data both input and output—that your model sees and gains knowledge from. fit the model on the training dataset, perform process initiates with providing high-quality data and subsequently training our machines (computers) by constructing machine learning models using the data and various algorithms.

Validation Data: Data validation ensures that the data is not tampered with by removing data mistakes from any project, giving the dataset accuracy, cleanliness, and completeness.

Testing Data: Testing data offers an objective assessment once our model has been fully trained. Our model will forecast some values when we feed in the testing data inputs (without seeing real outcome). Following prediction, we assess our model by contrasting its output with the real output found in the test data.

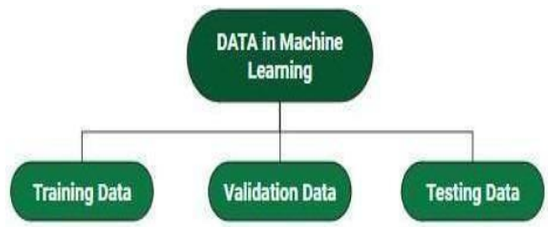

Fig. 1 Splitting the Data

1.1 Approach

The objective of this project is to develop a machine learning model capable of identifying fraudulent online payment transactions. XGBoost Classifier will be utilized for this purpose. To perform that it is very important to develop new ways or methods that can identify which are the things that make a payment fraudulent.

2. Materials and Methodology

2.1 Materials (Dataset Used):

Credit Card Fraud Dataset:

we need a dataset containing information about online payment fraud, so that we can understand what type of transactions lead to fraud. For this task, I collected a **dataset** from Kaggle, which contains historical information about fraudulent transactions which can be used to detect fraud in online payments.

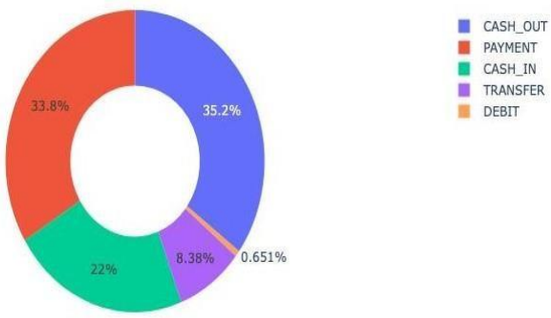

Fig. 2 Transaction Types

2.2 Methodology

2.2.1 Data Collection:

In the field of machine learning, data collection involves acquiring and then combining material that will be utilised in generating a model for carrying out tests as well as validation. It is the most important operation in machine learning pipeline because training data feeds into ML algorithm and results are fully depending on that with high probability.

2.2.2 Data Preprocessing:

In the context of machine learning, data processing is performed over a fixed piece or set to make it suitable for analysis and model training. Data cleaning, preprocessing and arranging is an essential step before using it into Machine Learning algorithms. Data preprocessing can significantly determine the success and productivity of machine learning models. Before building models, the imported dataset will be cleaned.

2.2.3 Analyzing Data and Visualization:

In machine learning, data analysis and visualization are repetitive processes that frequently require you to go back and improve your comprehension of the data as new information becomes available. Effective analysis and visualization help with improved feature engineering, model selection, and results communication to non-technical audiences.

2.2.4 Training and Testing:

A section of the dataset referred to as training data is used during model training in order to instruct the machine learning model on how to identify patterns and produce predictions. This dataset includes labels or target values for the output features in addition to input attribute values. However, test data is a separate subset of the dataset that is used to measure the performance of the model and its capacity to generalize to new, unobserved data points.

2.2.5 Predicting Output:

Predicting output is the process of utilizing a trained model to create predictions or forecasts based on input data. These predictions, which are basically the expected outcomes or answers produced by the model, might take a variety of shapes depending on the type of problem you're attempting to address.

2.2.6 XGBoost Classifier:

The objective of this project is to develop a machine learning model utilizing XGBoost (Extreme Gradient Boosting) for identifying fraudulent online payment transactions. XGBoost is a robust, scalable, and effective gradient boosting framework commonly utilized for a range of machine learning purposes, particularly for classification and regression tasks.

XGBoost works in prediction:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$$

where:

- $\hat{y}_i$ is the predicted probability of the i-th transaction being fraudulent.
- K is the number of trees in the ensemble.
- $f_k(x_i)$ is the output (log-odds of fraud) of the k-th tree for the i-th transaction.

Fig. 3 Model Prediction

a. Data Loading:

The dataset is loaded, and necessary preprocessing steps are implemented. This involves converting transaction types and fraud labels into numerical values.

b. Feature and Target Selection:

Key features (type, amount, oldbalanceOrg, newbalanceOrig) are chosen to predict the target variable (isFraud).

c. Data Split:

The data is divided into training and testing sets. A smaller portion of the data is utilized initially to expedite the training process for preliminary assessments.

d. XGBoost Model Training:

The XGBoost model is trained using the training dataset. It constructs a series of decision trees, with each subsequent tree aiming to enhance predictions based on the errors of the previous trees.

e. Prediction Generation:

The trained model is employed to forecast fraud on the test dataset. The collection of trees collaborates to form the final prediction by combining the outputs from all individual trees.

3. Result:

With a remarkable accuracy of 99.93%, the XGBoost Classifier model demonstrated remarkable performance on the wound healing prediction challenge. The model exhibits perfect prediction capabilities, as evidenced by its perfect F1 score, recall, and precision, all at 1.00. This indicates that there were neither false positives nor false negatives in the model's identification of all cases of healed and non-healed wounds. Due to its great performance, the model is very dependable for this classification test, indicating that it has successfully learned the distinctive features in the dataset.

XGBoost Classifier	
Accuracy	0.99933
Precision	1.00
Recall	1.00
F1 Score	1.00

Table1: Success rate table

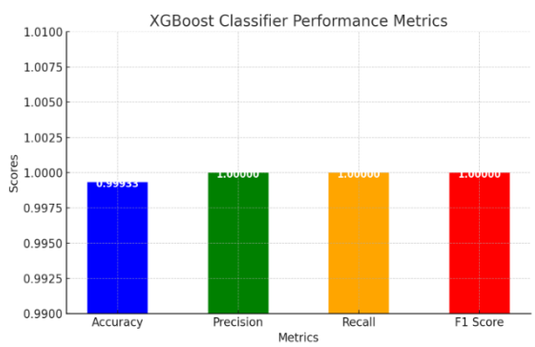

Fig. 4 Graphical representation of the performance metrics for the XGBoost Classifier.

4. Conclusion

The project for Identifying & educated consumer of online payment fraud detection used the XGBoost classifier to predict fraudulent transactions from a credit card transaction data. The model was pre-processed, feature selected and trained in XGBoost which performed well with high accuracy due to imbalanced data issues as it provided better insights through precision, recall and F1 score metrics. This indicates the generalization power of this model because it can predict a new transaction as fraudulent when we do not see during training. But, some are like finding the best of available features or adding anomaly detection than beyond this even more research has to be done were I can extend my knowledge. In sum, the project succeeded in constructing a machine learning model.

References:

1. Vanini, P., Rossi, S., Zvizdic, E., & Domenig, T. Online payment fraud: from anomaly detection to risk management. *Financial Innovation*, Vol.9, No.66, 2023, 1-13. https://doi.org/10.1186/s40854-023-00449-7.

2. Almazroi, A. A., & Ayub, N. Online payment fraud detection model using machine learning techniques. *IEEE Access*, Vol.11, 2023,137188-137203. https://doi.org/10.1109/ACCESS.2023.3339226.

3. Singh, J., & Kaur, P**.** Fraud detection in online transactions using machine learning. *Bournemouth University*, 2023. https://doi.org/10.13140/RG.2.2.29971.66088.

4. Chen, T., & Guestrin, C**.** XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference*,2016. https://doi.org/10.1145/2939672.2939785.