Improving the classification accuracy in the case of Unbalanced dataset

Konijeti Sri Vyshnavi
Department of Computer Science (Artificial Intelligence) Amrita Vishwa Vidhyapeetam
Amritapuri, Kerala, India
amenu4aie20042@am.students.amrita.edu

Maddala H S M Krishna Karthik
Department of Computer Science (Artificial Intelligence) Amrita Vishwa Vidhyapeetham
Amritapuri, Kerala, India
amenu4aie20046@am.students.amrita.edu

Methuku Samhitha
Department of Computer Science (Artificial Intelligence) Amrita Vishwa Vidhyapeetham
Amritapuri, Kerala, India
amenu4aie20049@am.students.amrita.edu

Suravarapu Ankith
Department of Computer Science (Artificial Intelligence) Amrita Vishwa Vidhyapeetham
Amritapuri, Kerala, India
amenu4aie20070@am.students.amrita.ed

***Abstract* — Fraud has risen dramatically as a result of advances in technology and global connections. Detection and prevention are two approaches for preventing fraud. The dataset used to detect credit card fraud is heavily skewed and different sorts of misclassification errors may incur different costs, so it's critical to keep track of them. Classification techniques have promise for detecting both fraudulent and non-fraudulent transactions. The major goal of this research is to improve the XGBoost (eXtreme Gradient Boosting) strategy by applying resampling approaches to handle a class imbalance in datasets. XGBoost is a popular machine learning model that is utilized in areas such as fraud detection.**

***Keywords*—Fraud detection, Xgboost.**

INTRODUCTION

The epidemic caused by the outbreak of coronavirus has not stopped until today. Many things have changed in this epidemic. As a result of the outbreak, oral bans were introduced in many lands. With Lockdown people have begun to switch to online shopping using traditional shopping methods. Due to the rise of online trading, the rise of online-based cheating is rapidly increasing each year. Risk assessment is an important process that you should avoid distortion of facts and as a result many banks and organizations were taken over. They have invested billions of dollars to avoid distorting facts and ensuring online trading. It is very difficult to make any physical exchange. In fact, these exchanges are handled by machines that are free of human interference. Credit Card fraud is an illegal activity performed by criminals for short-term gain. And this will be known to the users, after a few days the fraud has occurred and they will respond later by registering a complaint about the fraud they have gone through. These exchanges are made using external installments, such as CCAvenue, ICICI Payseal, and PayPal, and other false Mastercard exercises may be ordered from (I) Mastercards included (ii) taken or lost (iii) no (card holder data obtained) (iv) supermarket fraud fraud. The main purpose and objective of non-distorted identification distortions is to identify and ensure that certain transactions are made by a real client (cardholder) or others.

Credit Card Fraud is a very important way to find out if the work is done by the owner or any other person so by using XGBoost (eXtreme Gradient Boosting) we can detect fraudulent activity. XGBoost was developed under the Gradient Boosting structure and was created by Chen and Guestrin, which is intended to produce in depth, flexibility, and simplicity. The envisaged goal of support is to combine the advancement of weak class dividers with low accuracy in order to build a strong separator with the use of better planning.

XGBoost talks about Extreme Gradient Boosting. The nominees for the University of Washington are Tianqi Chen and Carlos Guestrin. XGBoost is one of the most important AI models you can work with with equal data and is an AI team method used to help Gradient. It is very different from other slopes that support the slope and is still an excellent choice for a variety of real AI stories and is an open library and used for accurate scaling to see the best tree model. Support shows a group-based reading calculation that transforms weak students into strong assessors.

1.Calculates XgBoost with steps in steps.

2.We have to take Data

3.To build a base leaner

4.We have to count the remnants

5.Build a Decision Tree and calculate the same weights

6.We should count Gain

7.Possible predictions

8.Start the multiplication from point 3 to 6 and at the end of the repetition, the remainder will be very small.

9.Experimental prediction in the repetition model has a small remnant.

LITERATURE REVIEW

[1] This paper is about Artificial Neural Network, Machine Learning, BackPropagation, Credit Card Fraud, Decreased Gradient. The demand for credit card payments is growing exponentially and is fraudulent. This model was created, using the ANN (Artificial Neural Network) algorithm and Backpropagation.

The strategy used to upgrade the ANN device and the Logistic Regression. This model also has low accuracy in the data sets that test the use of ANN (Artificial Neural Network) and BackPropagation methods, which provide 99.96% less accuracy than other algorithms.

This will give you an idea of each transaction and will detect fraudulent sales in real time.

Using a Credit Card Customer database with 31 attributes of our ANN model. The first 30 attributes have records associated with Customer Age, Name, Gender, etc. and the last attribute of both zero and 1 will give the effect of action.

PCA strives to reduce the variety of items within a database while holding “key components”, which are defined as scales, existing components are a line component designed to be impartial and to deal with large variations in data.

The scope of the future associated with this project is enormous. This model can then be sent to a financial institution gadget to detect fraud.

This is a proposed paper about -Credit card, fraud detection, type inequality, efficiency, XGBoost, hyperparameter, parameter adjustment.

XGBoost is a portable computer technology used in domains such as fraud detection and handling of the type of inequality that causes overcrowding if not handled properly.

[2] This is a proposed paper about -Credit card, fraud detection, class inequality, efficiency, XGBoost, hyperparameter, parameter adjustment.

XGBoost is a popular computer-assisted model used in domains such as fraud detection and dealing with class inequalities that create overcrowding when not handled properly.

This paper proposes to manage type inequality in data sets without the use of re-sampling techniques and an improved XGBoost (OXGBoost) strategy. In the proposed method, RandomizedSearchCV hyperparameter optimization is to determine the appropriate XGBoost parameters.

The most widely used techniques in the Bagging and Boosting ensemble. Bagging (Bootstrap Aggregation) is a meta algorithm designed using Breiman that creates multiple basic models in a sequence and re-sampling is completed with different educational information taken from real educational facts by line sampling by replacing and finally, the results are built-in. by a majority vote.

The best way to learn about Ensemble proven is to predict the category of binary split problems. The XGBoost mannequin using a tree-growing method with both extracts and especially the gradient gi and hessian independently creates a growing tree to deal with species inequality.

XGBoost has a series of general parameters, a booster, an expert function, and a command line. Paper on the impact of development on the XGBoost model to deal with class inequality in the database itself.

[3] This paper includes fraud detection, XGBoost, Focal loss, Weighted cross-entropy (W-CEL) loss, XGBoost inequality.

This leads to huge losses on both banks and customers. One of the biggest challenges is the difficulty of finding the features of credit card fraud.

Terms and contributions are provided through the author:

- Using unequal parameter φ from W-loss to Focal loss

- Using the tuning-hyperparameter to the value γ.

- Evaluating the Modified Focal Loss on XGBoost Imbalance manipulation of comparison metrics: accuracy, memory, and MCC

This article contains

Section 1 on the values of this study,

The second section deals with the statistics used,

Section 3 discusses how to look and act,

Section 4 which deals with tests and results in addition, as well

Section 5 which deals with the observation made.

XGBoost is a popular algorithm that accepts an additional management system with a second order, a 1-order gradient, and a second product order called hessian for job losses.

XGBoost parameter, as alternatively the parameters γ and α parameters are calculated using GridSearchCV from scikit learn.

In this article, we have proposed a Focal Loss Solution to resolve the unequal data issue without any prior processing measures such as sampling and exclusion.

[4] In a paper named XGBoost Advanced Model Based on Spark Credit Card Prediction Hongwei Chen, He Ai, Dawei Dong, and Weiwei Yang and discuss credit card fraud and unequal data using XGBoost models, Spote, Smote . The Smooth Algorithm process is also explained step by step and has a positive impact on the data treatment of the problem of inequality. And the next XGBoost Algorithm is an integrated learning algorithm based on the Development strategy. In this way there are two types: cuttings for new students and cuttings for Strong students.

In this way they describe the step-by-step process:

- Build a CAET decision tree.

- Calculate the static gradient for each sample

- Discover a new tree by doing selfish algorithms and gradient calculations.

- The correct point for the division of the new decision tree.

- They finally did these steps until a final model was found

Spark is a large data-based computer platform open to UC Berkeley that solves problems such as map reduction and, ultimately, concludes that data inequality and large amounts of data in the field of credit card fraud and improved the XGBoost model. based on Spark. They also combine both the Smote algorithm and the XGBoost algorithm. And it works with good output. And especially to speed up the Spark environment.

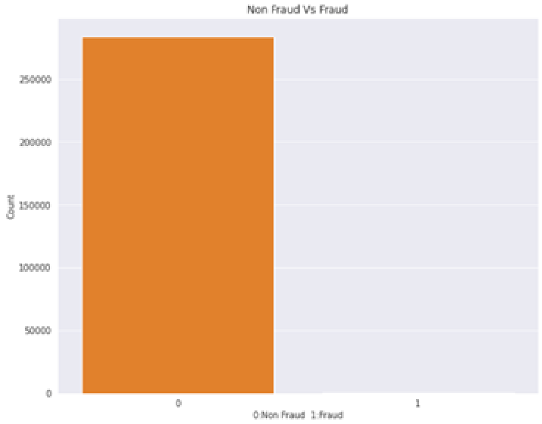
DATASET

This study used the "Credit Card Fraud Detection" dataset, which is a normal imbalanced machine learning dataset. This dataset includes 492 frauds out of 284,807 credit card transactions done by European cardholders over two days in September 2013. Frauds account for 0.172 percent of all transactions, hence the sample is severely skewed. It only accepts numeric input variables that have been transformed using PCA. The original features and background information about the data are not provided due to data confidentiality concerns.

V1, V2, V3, V4,...V27, V28 features are the primary components derived using PCA; the only attributes not altered by PCA are 'Time' and 'Amount.' The feature 'Time' stores the number of seconds that have passed between each transaction and the first transaction in the dataset. The 'Amount' feature represents the transaction amount. 'Class' is the answer variable, and it has a value of 0 when there is no fraud and 0 when there is a fraud transaction.

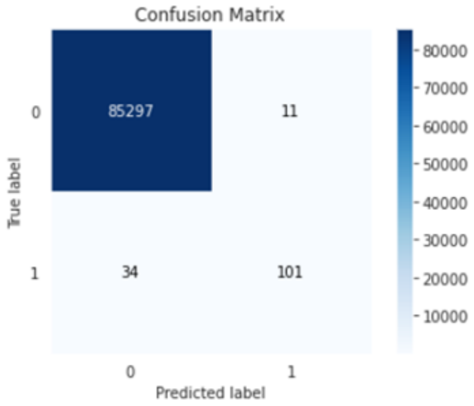
METHOD

Initially, import necessary libraries like numpy, pandas, xgboost etc. And then read the credit card fraud detection dataset and explore the dataset by checking the existence of null in any column, finding the variable type in each column and finding the statistical information about each variable and each column. Then visualize the data distribution using seaborn and matplotlib libraries as follow.



We can observe that out of 284,807 transactions, 284,315 (99.83 percent) were classified as normal, while only 492 were classified as fraudulent (0.17 percent). Despite the fact that fraud transactions are tiny, they can add up to a large expense, resulting in billions of dollars in lost revenue each year.

Then, using the train test split function, we separated the input variables from the target variables and divided the data into train and test sets. The train test split function divides data into train and test sets using a randomizer. 70% of the data for training and 30% of the data for testing were defined in this scenario. The random seed is used to ensure that all runs use the same data. Now, in the credit card fraud detection dataset, develop a simple XGBoost model capable of recognizing whether a transaction is fraudulent or not. Also, the performance of the model that has been constructed will be examined.

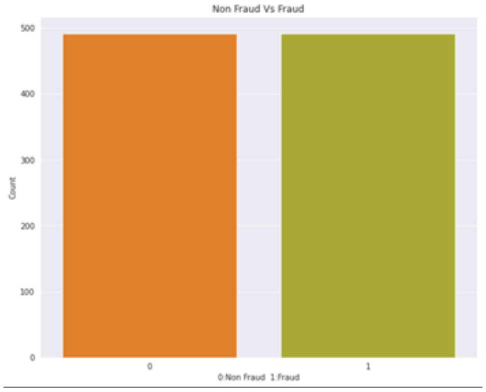


The model's accuracy is 99 percent. Although the accuracy was outstanding, the model incorrectly identified certain fraudulent transactions when looking at the confusion matrix. In a severely unbalanced dataset, accuracy does not indicate a correct metric for a model's efficiency. Before considering any performance evaluation measures, there should be a way for balancing the data.

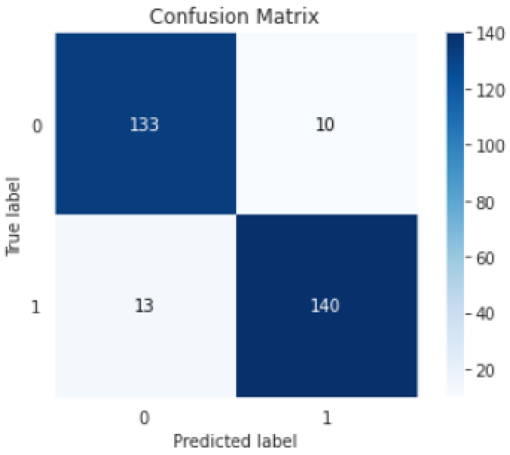
Methods such as undersampling and oversampling approaches are commonly used to balance skewed datasets. The algorithm becomes more sensitive to fraudulent transactions when the sample is changed. The practice of deleting significant class records from a sample is known as undersampling. In this situation, random records from the class (No fraud) must be removed in order to acquire a number of records similar to the quantity of the minority class (fraud) needed to train the model. Oversampling is the polar opposite of undersampling; it entails including minority class records (fraud) in our training sample, hence raising the proportion of fraud records overall. There are strategies for obtaining samples from the minority class, such as copying existing records or creating new ones artificially.

To obtain a uniform divide between fraud and genuine transactions in the current model, we will apply the undersampling technique. This will result in a useful classifier.

Following is the data distribution after undersampling.



In addition, the new model's performance for balanced data will be examined.



The model is 92 percent accurate. Although accuracy has dropped, sensitivity has risen dramatically. We can see a substantially higher percentage of correct classifications of fake data in the confusion matrix. Unfortunately, a higher number of fraud classifications usually always equals a higher proportion of legitimate transactions that are also categorized as fraudulent.

We'll now evaluate the model's performance using the new model for the original dataset.

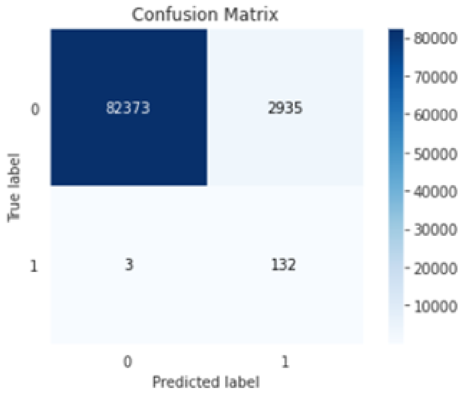
The model is quite accurate. We'll also use the ROC and AUC curves to assess the classifier's performance. The ROC curve is a probability curve that displays how well a classifier can distinguish between two items based on two parameters. Those are the number of times the classifier got it right and the number of times the classifier got it wrong. The AUC represents the degree or measure of separability and is generated from the ROC curve. The AUC calculates the area under the curve to describe the ROC curve in a single value.

The AUC indicates how well the model predicts 0s as 0s and 1s as 1s. The AUC value might be anywhere between 0.0 and 1.0. In this scenario, the higher the AUC, the better the model is at detecting fraudulent and valid transactions.

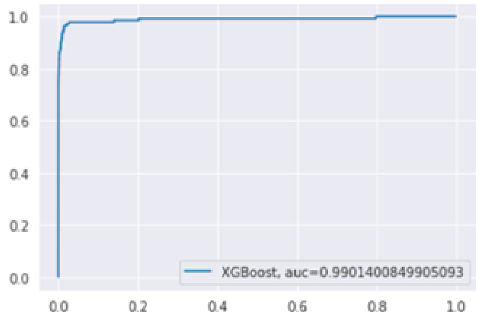
Programming language: Python

CONCLUSION

The classifier performance is measured and it has an accuracy of 96%. In connection to the unique approach, we have accelerated from 75 percent to 96 percent of accurately identified jobs, achieving a very first-class quantity in detecting fraud. In retrospect, the overall performance of a well-defined ordinary operation decreased from 99% to 96%.



We can see a substantially higher percentage of correct classifications of fake data in the confusion matrix. The ROC and AUC curves are also used to assess the performance of the classifier.



As we all know, the greater the AUC, the better the model is at distinguishing between legitimate and fraudulent transactions. With an AUC of 0.99, we can observe that the model has a pretty good result.
Remember that we want to determine if this exchange is appropriate. In general, the value of losing a fraudulent job is often greater than by chance keeping apart a desirable job such as fraud. Finding a stable coaching paradigm and moving forward accordingly is one of the problems.
As a way to further improve model performance, there are a variety of bendy input test methods, performing unique "Pre-Data Processing" and "Feature Engineering" techniques.

FUTURE WORK

The scope of the future is too great to relate to this project. There are many other Re-sampling methods to stabilize a data set and various computing devices that receive strategic information are used to detect the effects of credit card fraud effectively. This mannequin can then be sent to a bank machine to detect fraud. Banks can install this mannequin where and when the fraudulent activity is in progress then this mannequin will detect fraud almost with the correct accuracy. Delivery will be done using Cloud Services. This will help banks to detect fraud quickly and without interference. Several activities are listed that will be performed in the next step of this experimental study. We will develop a model of the difficulty of class inequality in order to find a balance between sensitivity and accuracy.

REFERENCE

[1] S. C. DUBEY, K. S. MUNDHE AND A. A. KADAM, "CREDIT CARD FRAUD DETECTION USING ARTIFICIAL NEURAL NETWORK AND BACKPROPAGATION," 2020 4TH INTERNATIONAL CONFERENCE ON INTELLIGENT COMPUTING AND CONTROL SYSTEMS (ICICCS), 2020, PP. 268-273, DOI: 10.1109/ICICCS48265.2020.9120957.

[2] C. V. Priscilla and D. P. Prabha, "Influence of Optimizing XGBoost to handle Class Imbalance in Credit Card Fraud Detection," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), 2020, pp. 1309-1315, doi: 10.1109/ICSSIT48917.2020.921420

[3] TRISANTO, D., RISMAWATI, N., MULYA, M.F. AND KURNIADI, F.I., 2021. MODIFIED FOCAL LOSS IN IMBALANCED XGBOOST FOR CREDIT CARD FRAUD DETECTION. INT J INTELL ENG SYST, 14, PP.350-8.

[4] H. Chen, H. Ai, Z. Yang, W. Yang, Z. Ye and D. Dong, "An Improved XGBoost Model Based on Spark for Credit Card Fraud Prediction," *2020 IEEE 5th International Symposium on Smart and Wireless Systems within the Conferences on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS-SWS)*, 2020, pp. 1-6, doi: 10.1109/IDAACS-SWS50031.2020.9297058.

[5] Science, WARSE The World Academy of Research in, and Engineering. "Credit Card Fraud Detection Using Imbalance Resampling Method with Feature Selection." International Journal of Advanced Trends in Computer Science and Engineering , 2021.

[6] C. Whitrow, D. Hand, J. Juszczak, D. Weston, and N. M. Adams, "Transaction aggregation as a strategy for credit card fraud detection," Data Mining and Knowledge Discovery, pp. 30–55, 2009

[7] Nishant Sharma, "CREDIT CARD FRAUD DETECTION PREDICTIVE MODELING", A Paper Submitted to the Graduate Faculty of the North Dakota State University of Agriculture and Applied Science, May 2019

[8] K. Elissa, "Title of paper if known," unpublished.

[9] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[10] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[11] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.