Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 6- Issue 1, Paper 30,January 2023

Fraud Detection in Financial Transaction using Synthetic Minority Oversampling and CatBoost Model

Dr. M. Nirmala Devi[1], A. Priyankaa[2], S. Roshini[3], R. Mahalakshmi[4]
Department of Computer Science and Engineering
Thiagarajar College of Engineering, Madurai, India
roshinis@student.tce.edu

***Abstract*—The number of online transactions has skyrocketed over the last few decades. This leads to an increase in financial fraud transactions, resulting in financial losses. It's critical to develop trustworthy fraud detection systems for fraud prevention. This research introduces a fraud detection machine learning algorithm based on CatBoost. Three oversampling methodologies are employed to handle the unbalanced dataset: Synthetic minority oversampling technique (SMOTE), RandomOverSampler and adaptive synthetic (ADASYN) to generate a balanced dataset. Comprehensive classification measurements are used, which include fundamental, combined, and graphical measurements to evaluate and compare the performance of these models. As a result, Synthetic Minority Oversampling Technique (SMOTE) has been found to give more accurate outcomes than the other two resampling procedures. The CatBoost method is then applied to the balanced dataset, and the testing results demonstrate that the CatBoost-based model has an ideal PR AUC score of 99.5% and accuracy of 99.8%.**

Keywords— Synthetic Minority Oversampling Technique (SMOTE), Adaptive Synthetic Minority Oversampling Technique, RandomOverSampler (ADASYN), CatBoost, Fraudulent Transaction.

I. INTRODUCTION

Fraudulent actions have been on the rise in a variety of businesses around the world, particularly in the financial sector. Credit card fraud is the most serious in financial institutions, and it must be prevented as quickly as possible. According to PwC's global economic crime and fraud survey, 46% of surveyed organizations have experienced fraud in the past 24 months. Fraud detection approaches must investigate and strictly handle financial fraud in order to drastically limit the effects. Financial fraud is difficult to detect due to a lack of information discovery and deep understanding into the type or patterns of transactions occurring and their tendencies.

Machine learning, a popular approach for extracting information from massive datasets, is a popular tool for detecting and avoiding financial fraud. Support Vector Machine, Naive Bayes, Logistic Regression, K-Nearest Neighbor method, Random Forests, Data Mining, Light Gradient Boosting Machine, and other advanced systems have been proposed to solve this problem. Although these advanced systems have achieved considerable performance, they are prone to fail in complex situations and limited to detection speed. As a result, the development of fraud detection technologies is critical.

The CatBoost model, which employs a large number of category characteristics, is much faster to process than other methods like XGBoost or Random Forest. CatBoost has built-in parameters to reduce overfitting as Overfitting can easily happen in boosting algorithms because they are tree-based algorithms.

Machine learning algorithms, on the other hand, operate best when there is a balanced distribution of classes. Various treatments have been investigated in the previous decades to address the issue of skewed datasets. Three types of solutions are typically proposed in these studies: data-level, algorithm-level, and ensemble solutions.

To generate a balanced dataset, three resampling strategies are used in this paper: synthetic minority oversampling technique (SMOTE), adaptive synthetic (ADASYN), and RandomOverSampler. The balanced dataset acquired is trained using the CatBoost machine learning (ML) technique. The optimal case for detecting financial fraud is determined by comparing the performance of machine learning algorithms based on three resampling strategies.

II. RELATED WORKS

The rapid rise of e-commerce and online-based payment options has coincided with an empirical universe of economic fraud, in which credit card fraud has been more difficult to avoid for many years; various researchers have created a number of data mining-based approaches to combat this problem. There has recently been a lot of buzz about using machine learning algorithms instead of data mining approaches to detect credit card fraud [13]. Implementation of efficient fraud detection algorithms employing machine-learning approaches is critical in assisting fraud investigators [8].

SMOTE has inspired several approaches to counter the issue of class imbalance, and has also significantly contributed to new supervised learning paradigms, including multilabel classification, incremental learning, semi-supervised learning, multi-instance learning, among others. It is a standard benchmark for learning from imbalanced data [2]. When the data is substantially skewed, the machine learning model has a hard time detecting fraudulent transactions. Synthetic Minority Oversampling Technique (SMOTE) is effective in increasing the performance of unbalanced data classification when applied on machine learning algorithms [5]. The ability to forecast positive classes improved significantly after utilizing the SMOTE based Oversampling technique, which is a data-point approach [4]. According to the results, the random forest and decision tree algorithms performed best for fraud detection.

Two resampling approaches - (SMOTE) and adaptive synthetic (ADASYN) are used to handle an imbalanced dataset to obtain the balanced dataset [6]. The ML methods presented demonstrate positive results of categorization for fraudulent activity after resampling the dataset.

A critical modification to Synthetic Minority Oversampling Technique (SMOTE) is proposed for highly imbalanced datasets, where the generation of new synthetic samples are directed closer to the minority than the majority [15]. In this way, the line of distinction between the two classes will be precisely defined and all data samples will be positioned inside their class borders to establish accurate prediction of the classifiers established.

For the detection of credit card frauds, three boosting techniques are used: CatBoost, XGBoost, and Stochastic gradient boosting algorithms [1]. When compared to XGBoost and SGB boosting algorithms for the categorization of fraudulent or non-fraudulent transactions, CatBoost method has the best evaluation of metric parameters such as precision, recall, and confusion matrix.

For fraud detection, a machine learning approach based on CatBoost has been developed [9]. Feature engineering is used to develop extremely important features and feed them into CatBoost for classification to improve detection accuracy. CatBoost is a good option for Big Data machine learning implementations [3]. CatBoost is well-suited to categorical, heterogeneous data machine learning problems.

III. METHODOLOGY

The dataset for the project was gathered from Kaggle (Synthetic Financial Datasets For Fraud Detection). It consists of nearly 6 million rows. The data is visualized after Data Cleaning and Feature Engineering. The dataset is imbalanced between the two labels : fraud and non-fraud. The imbalanced data do not show good performances to detect the fraudulent activities, so there is a need to resample this dataset to a balanced dataset. Resampling techniques - SMOTE and ADASYN show the positive results of classification for fraudulent activities. So

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 6- Issue 1, Paper 30,January 2023

Oversampling methods such as SMOTE, ADASYN and RandomOverSampler are applied to get a balanced dataset. Comparing these oversampling methods, SMOTE gave better results. Also SMOTE gives higher accuracy for different ML models. So SMOTE is applied for resampling our dataset.

A. SMOTE Oversampling Technique

Oversampling the minority class examples is one technique to balance the data. Before fitting a model, this can be performed by simply duplicating minority class samples in the training dataset. This can help to balance the class distribution, but it doesn't give the model any extra information.

Synthesizing new instances from the minority class is an improvement over replicating examples from the minority class.The Synthetic Minority Oversampling Technique(SMOTE) is perhaps the most extensively used method for creating new samples. SMOTE works by picking close-together instances in the feature space, drawing a line in the feature space between the examples, and drawing a new sample along that line.

SMOTE picks a minority class instance at random and searches for its k closest minority class neighbors. The synthetic instance is then constructed by randomly selecting one of the k nearest neighbors b and connecting a and b in the feature space to form a line segment. The Synthetic instances are made by convexly merging the two selected examples a and b.

The SMOTE Algorithm includes the following steps:

Step 1: Let the minority class set be . For each ∈ , the k-nearest neighbors of are obtained by computing the Euclidean distance between and every other sample in set .

Step 2: According to the imbalanced proportion, the sampling rate is set . For each ∈ , the set $_1$ is constructed by randomly selecting examples (i.e. $_1$, $_2$…) from its k-nearest neighbors.

Step 3: For each example ∈ $_1$($=1,2,3..$), the formula used to generate a new example is,

$$' = + (0,1) * | - |,$$

where $(0,1)$ denotes a number between 0 and 1 at random.

B. CatBoost Model

After achieving a balanced dataset, CatBoost(machine learning technique) is applied to the dataset to develop a Fraud Detection Model.

CatBoost, is a sophisticated machine learning technique that provides state-of-the-art outcomes in a number of practical problems by implementing gradient boosting and employing binary decision trees as base predictors. CatBoost, unlike other gradient boosting implementations such as eXtreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM), uses a novel ordering strategy to address prediction shifts induced by a specific type of target leakage.

Assume you have a predictive model F. It is obtained after several steps of boosting based on the goals of all training examples, which causes a shift in the distribution of F (X_k) for a training example. For a test case X, take X_k from the F (X)|X distribution. The learned model's forecast shifts as a result of this. To eliminate prediction shifts, the CatBoost uses ordered boosting, which is a variation of the basic gradient boosting approach.

In ordered boosting, a tree is trained on a subset of the data set and then used to calculate residuals for a subset it hasn't seen yet. CatBoost accomplishes this by permuting the data at random. Specifically, the dataset is permuted at random, and the average label value is computed for each sample with the same category value placed before the provided one in the permutation. To improve the algorithm's robustness, several permutations are applied on the

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 6- Issue 1, Paper 30,January 2023

training dataset. On the basis of a random permutation, gradients are sampled. These are the same permutations that are used to calculate categorical feature statistics. Different permutations are used to train different models. As a result, several permutations do not result in overfitting. In the permutation $\square = (\square_1, \ldots . \square_{\square})$, $\square_{\square\square,\square}$ is replaced with the equation,

$$\frac{\Sigma_{j=1}^{p-1} [x_{\sigma_j,k} = x_{\sigma_p,k}] Y_{\sigma_j} + a \cdot P}{\Sigma_{j=1}^{p-1} [x_{\sigma_j,k} = x_{\sigma_p,k}] + a},$$

which additionally includes a prior value $\square$ and a parameter $\square > 0$, which is the prior's weight.

CatBoost employs the Target-Based with Prior ordering concept, in which the values for each example are determined only by the observed history. As a result, the accuracy of CatBoost will be better for data with categorical attributes.

IV.PROPOSED APPROACH

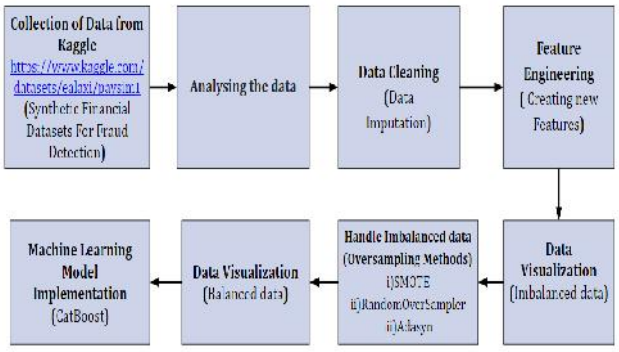


Figure 1. Flow Diagram

A. Analyzing the data

This is the quickest approach to learn more about the dataset. In this dataset, it is found that fraud occurs in only two of the five types of transactions: 'TRANSFER', in which money is sent to a customer fraudster, and 'CASH OUT', in which money is paid to a merchant who pays the customer fraudster in cash. Because neither the nameOrig nor the nameDest features encode merchant accounts in the expected manner, they are removed from the data. The feature isFraud is always set when isFlaggedFraud is set, because isFlaggedFraud is only set 16 times in an apparently meaningless fashion, this feature may get ignored and can be eliminated from the dataset without losing information.

B. Data Cleaning

It is found that fraud only happens in 'TRANSFER's and 'CASH OUT's after analyzing the data. As a result, the

related data is solely analyzed. The columns 'nameOrig','nameDest', and 'isFlaggedFraud' are removed from the EDA because they are not useful for analysis.

The original dataset's total number of records is,
Total Records = 6362620,
Majority Class (isFraud=0) = 6354407(99.7 %) , Minority Class (isFraud=1) = 8213 (0.3 %).

Total number of records obtained after data cleaning,
Total Records = 2770409 ,
Majority Class (isFraud=0) = 2762196 (99.7 %) , Minority Class (isFraud=1) = 8213 (0.3 %).

Although the transacted 'amount' is non-zero, the fraction of fraudulent transactions with 'oldBalanceDest' = 'newBalanceDest' = 0 is 0.4955558261293072.

0.0006176245277308345 is the fraction of legitimate transactions with 'oldBalanceDest' = newBalanceDest' = 0 but a non-zero transacted 'amount'.

C. Data Imputation

The value of 0 is replaced with -1 in oldBalanceDest and newBalanceDest since the destination account balances being zero is a significant indicator of fraud.
If (x.oldbalanceDest == 0) & (x.newbalanceDest == 0) & (x.amount != 0),
then ['oldbalanceDest', 'newbalanceDest']] = - 1

There are multiple transactions in the data that have zero balances in the originating account before and after a non-zero amount is exchanged. As a result, a null value for the value 0 has been substituted.
If (x.oldbalanceOrg == 0) & (x.newbalanceOrig == 0) & (x.amount != 0),
then ['oldbalanceOrg', 'newbalanceOrig']] = np.nan

D. Feature Engineering

When the values of oldbalanceOrg, newbalanceOrig, oldbalanceDest,newbalanceDest are zero, when the amount is not 0, there is a good chance that the transaction is fraudulent. So two new features are created (columns) for recording errors in the originating accounts, which combine oldbalanceOrg, newbalanceOrig, amount, and destination accounts, which combine oldbalanceDest, newbalanceDest, amount for each transaction. These new features prove to be crucial in generating the best results from the machine learning method which will be ultimately utilized.

data['errorBalanceOrig'] = data.newbalanceOrig + data.amount - data.oldbalanceOrg

data['errorBalanceDest'] = data.oldbalanceDest + data.amount - data.newbalanceDest

E. Data Visualization

The best method to ensure that the data contains adequate information is to try to immediately visualize the distinctions between fraudulent and genuine transactions.

The target label is extremely imbalanced, as seen in Figure 2., with only 0.129 percent of fraudulent data, which is insufficient for machines to learn and detect fraud when fraud transactions occur.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 6- Issue 1, Paper 30,January 2023

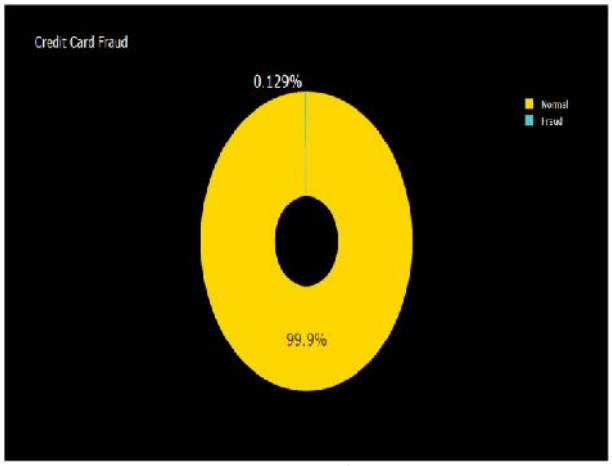


Figure 2. Data visualization for imbalanced data

Hence Oversampling or undersampling should be considered as in the case of the Financial Fraud dataset, severe imbalance is observed.

F. Pre-process imbalance data

An unequal distribution of classes within a dataset is referred to as data imbalance in Machine Learning. This problem occurs most frequently in classification jobs where the distribution of classes or labels in a dataset is not uniform. The unbalanced data, which appears to be large, is depicted in Figure 3.

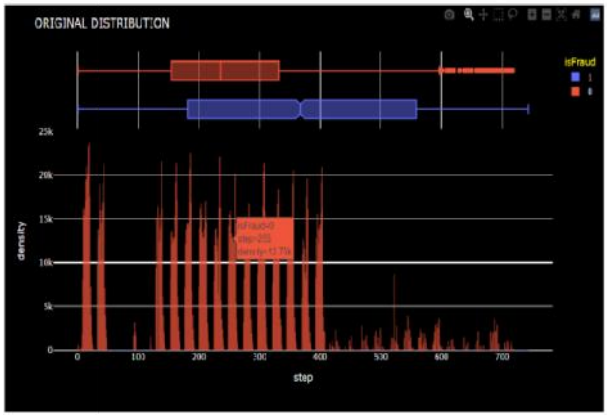


Figure 3. Distribution of imbalanced data

The resampling approach, which involves adding records to the minority class or deleting records from the majority class, is a simple way to handle this problem. One of the important conclusions proposed in is that oversampling outperforms undersampling for many classifiers and results in greater scores in several assessment criteria. To balance data, the dataset is oversampled.

G. Oversampling technique

Oversampling is the process of selecting respondents so that certain groups make up a higher percentage of the

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 6- Issue 1, Paper 30,January 2023

survey sample than they do in the population. It is the process of randomly picking and replacing instances from the minority class and adding them to the training dataset. It preserves all members from the minority and majority classes, so that no information from the original training set is lost. The dataset is subjected to three oversampling techniques: RandomOverSampler, SMOTE and ADASYN to determine the optimal technique.

1) RandomOverSampler : RandomOverSampler is a class that allows you to do random oversampling. By selecting samples at random with replacement, it over-samples the minority class(es).

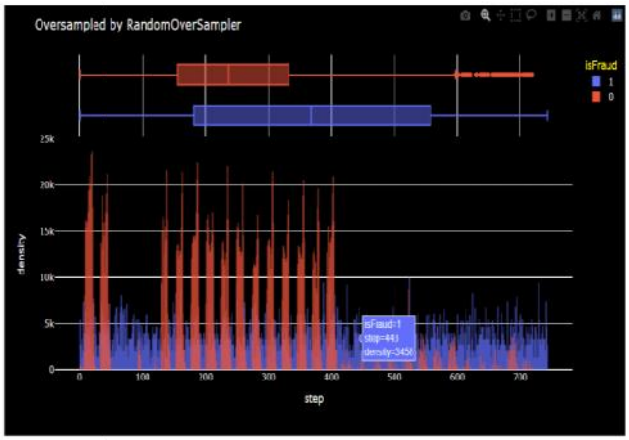


Figure 5. Oversampling using RandomOverSampler

From the following Figure 5, it can be identified that the dataset is prone to overfitting because in RandomOverSampler, the same information is replicated.

*2)Synthetic Minority Oversampling Technique (SMOTE***) :** This method is a statistical technique for evenly increasing the number of cases in our dataset. The component generates new instances based on current minority situations which are provided as input.

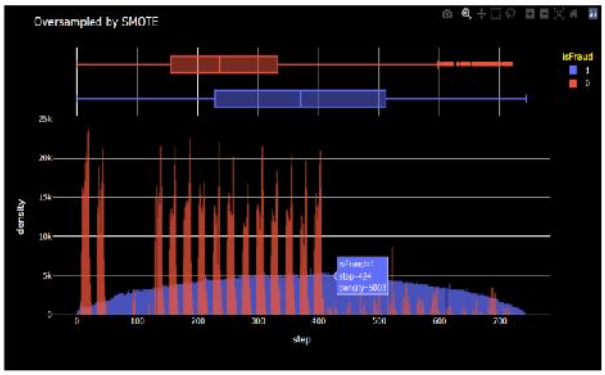
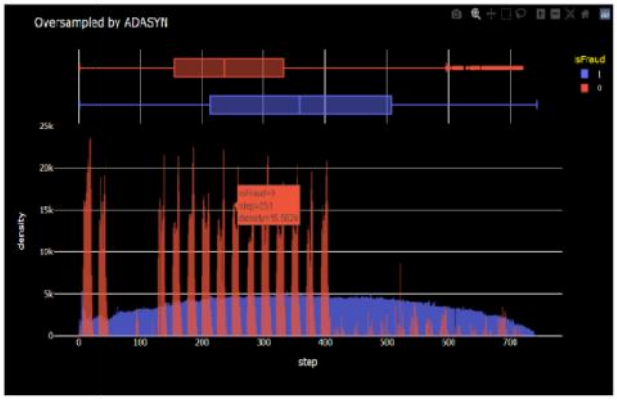


Figure 4. Oversampling using SMOTE

From Figure 4, It is observed that in the distribution of the created target data, minority class(isFraud=1) is spread over a wide range.

3) Adaptive Synthetic Minority Oversampling Technique (ADASYN) : This method is similar to SMOTE, but it creates a variable number of samples based on an estimate of the local distribution of the class to be oversampled.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 6- Issue 1, Paper 30,January 2023



A comparatively larger number of synthetic data is created in regions of a low density of minority class than higher density regions as depicted in Figure 6. Thus, the minority class is sparsely distributed.

H) Selection of best Oversampling Technique

Total Records = 2770409
Majority Class = 2762196 (99.7 %)
Minority Class = 8213 (0.3 %)

Table I. Majority and Minority Class Samples after applying Oversampling Techniques

Oversampling Techniques	**Majority Class (isFraud = 0)**	**Minority Class (isFraud = 1)**
RandomOverSampler	2762196	2762196
ADASYN	2762196	2762670
SMOTE	2762196	2762196

From Table I, it is observed that after applying RandomOverSampler and SMOTE to the dataset, minority and majority class samples are equally balanced. But RandomOverSampler causes overfitting as the minority class samples are replicated. On the other hand, SMOTE generates new synthetic data using the k-nearest neighbor technique to balance the dataset. When ADASYN is applied to the dataset, the minority and majority class samples are merely balanced. In comparison to ADASYN and RandomOverSampler, SMOTE appears to be more efficient. Hence, SMOTE is chosen for balancing the dataset to achieve better accuracy in the ML model.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 6- Issue 1, Paper 30,January 2023

I) Balanced data

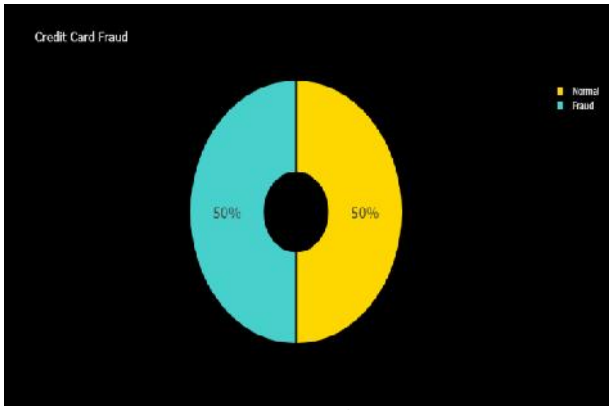


Figure 7. Data visualization for imbalanced data

From the Figure. 7, it is inferred that the data is balanced.

J) Model Implementation

CatBoost is well-suited to machine learning tasks which involve categorical, heterogeneous data. When compared to other boosting approaches, its training speed is estimated to be 4 times quicker on larger datasets and 2 times faster on smaller datasets. As this dataset consists of nearly 6 million rows, the CatBoost model has been chosen. SMOTE technique is used for preprocessing and the CatBoost model for fraud detection.

The most typical criterion for evaluating classification performance is Accuracy. The metric - Accuracy can't be used because the dataset is highly imbalanced. Just classifying all observations as the majority class yields a high accuracy score. When the data is significantly imbalanced, PR AUC is excessively focused, because PR AUC is more concerned with the positive class (PPV and TPR) and less concerned with the frequent negative class. Hence, Recall, precision, F-1 Score are the recommended evaluation criteria. The F-1 Score is used to assess minority class categorization in imbalanced classes.

Accuracy score = $\frac{(\square\square+\square\square)}{(\square\square+\square\square+\square\square+\square\square)}$

Precision = $\frac{\square\square}{(\square\square+\square\square)}$

Recall = $\frac{\square\square}{(\square\ +\square\square)}$

F1-score = $\frac{2*(\square\square\square\square\ \ \ \square*\square\square\square\ \ \ )}{(\square\square\square\square\ \ \ +\square\square\square\ \ \ )}$

False Positive Rate = $\frac{\square\square}{(\square\ +\square\square)}$

Table II. Evaluation Metrics

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 6- Issue 1, Paper 30,January 2023

CatBoost Model Implementation	ROC AUC	PR AUC	F1 Score	Precision
Before balancing the data	0.977	0.804	0.894	0.841
After balancing the data (SMOTE)	0.997	0.995	0.997	0.995

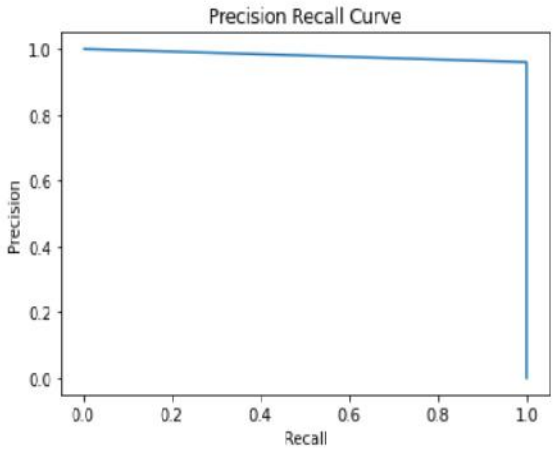


Figure 8. Before balancing the da

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 6- Issue 1, Paper 30,January 2023

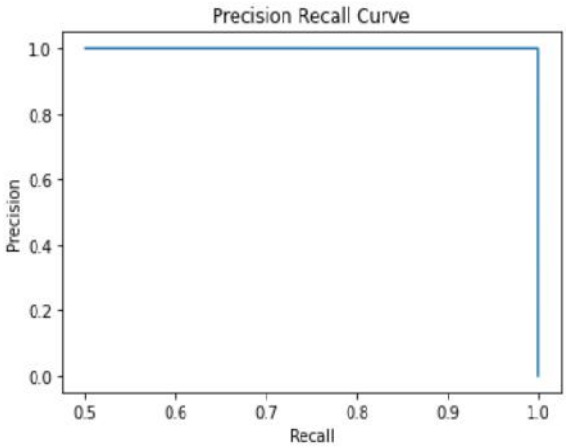


Figure 9. After balancing the data (SMOTE)

From Table II.,it is noted that after balancing the dataset by applying the SMOTE algorithm in the CatBoost model, the PR AUC increases from 0.80 to 0.99.

The f1-score for neural networks (SMOTE) increased from 69.8 percent to 85.1 percent, f1-score for Naive Bayes (SMOTE) increased from 67.9 percent to 94.5 percent, f1-score random forest (SMOTE) increased from 69.8 percent to 94.3 percent, and the f1-score for the decision tree (SMOTE) increased from 56.8 percent to 91.2 percent [5]. The AllK-Nearest Neighbors (AllKNN) undersampling approach is combined with the CatBoost (AllKNN-CatBoost) model. With an AUC of 97.94%, a Recall of 95.91%, and an F1-Score of 87.40%, the new model beats earlier models [7]. The performance evaluation of the CatBoost model reaches the highest classification accuracy of 98.3% and AUC-ROC score of 97.1% [9].

The approach provided here, on the other hand, is capable of increasing the F1-Score value from 89.4% to 99.7%,PR AUC score from 84.1% to 99.5%, ROC AUC score from 97.7% to 99.7% and gaining an accuracy of 99.8%.

V. CONCLUSION

Machine learning methods are used to examine the fraud dataset in this work. This dataset is subjected to the processing data stage in order to improve the effectiveness of machine learning models. An effective resampling technique - SMOTE is used to achieve balanced data because this dataset is highly skewed. This dataset is subjected to the data processing stage in order to improve the efficacy and analyze the financial industry's fraud detection challenges, and it provides a machine learning solution based on CatBoost to improve detection efficiency.

We intend to continue exploring information and in-depth insight into the nature or patterns of financial transactions in the future, which will aid us in developing an inventive, efficient fraud detection model.

REFERENCES

[1]Sai Tejasri Yerramsetti, Prathima Gamini, Gayathri Devi Darapu, Vamsi Kaladhar Pentakoti, Vegesna Prudhvi Raju. “Detection of Credit Card Fraudulent Transactions using Boosting Algorithms,” *Journal of Emerging Technologies and Innovative Research (JETIR).* 2021; vol. 8; pp. 2-4.

[2] Alberto Fernandez, Salvador Garcia, Francisco Herrera, Nitesh V. Chawla. “SMOTE for Learning from Imbalanced Data: Progress and Challenges,” *Journal of Artificial Intelligence Research,* 2018; vol.61, pp. 863-905.

[3] John T. Hancock and Taghi M. Khoshgoftaar. “CatBoost for big data: an interdisciplinary review,” *Journal of Big Data.* 2020; vol. 7; pp. 1-45.

[4] Nhlakanipho Mqadi, Nalindren Naicker, Timothy Adeliyi. "A SMOTE based oversampling data-point approach to solving the credit card data imbalance problem in financial fraud detection," *International Journal of Computing and Digital Systems*. 2021; vol. 10; pp. 277-286.

[5] Adi Saputra and Suharjito. "Fraud Detection using Machine Learning in e-Commerce," *International Journal of Advanced Computer Science and Applications(IJACSA)*. 2019; vol. 10; pp. 332-337.

[6] S. Manlangit, S. Azam, B. Shanmugam, and A. Karim, Novel "Machine learning approach for analyzing anonymous credit card fraud patterns," *International Journal of Electronic Commerce Studies.* 2019; vol. 10; pp. 175-202.

[7] Noor Saleh Alfaiz and Suliman Mohamed Fati. "Enhanced Credit Card Fraud Detection Model Using Machine Learning," *Journal of Big Data*. 2022; pp. 11-17.

[8] Md. Noor Alam, Prajoy Podder(B) , Subrato Bharati, and M. Rubaiyat Hossain Mondal. "Effective Machine Learning Approaches for Credit Card Fraud Detection," *International Conference on Innovations in Bio-Inspired Computing and Applications*. 2020; vol. 1372, pp. 154–163.

[9] Yeming Chen and Xinyuan Han. "CatBoost for Fraud Detection in Financial Transactions," *IEEE International Conference on Consumer Electronics and Computer Engineering.* 2021; pp. 176-179.

[10] Doaa Al mhaithawi, Assef Jafar and Mohamad Aljnidi. "Example dependent cost sensitive credit cards fraud detection usingSMOTE and Bayes minimum risk," *SN Applied Sciences.* 2020; pp. 2-6.

[11] Pedro Silva, Catarina Macas, Evgheni Polisciuc, Penousal Machado. "Visualization Tool to Support Fraud Detection," *25th International Conference Information Visualisation.* 2021; pp. 3-7.

[12] Abdullahi A. Ibrahim , Raheem L. Ridwan , Muhammed M. Muhammed , Rabiat O. Abdulaziz , Ganiyu A. Saheed. "Comparison of the CatBoost Classifier with other Machine Learning Methods," International Journal of Advanced Computer Science and Applications. 2020; pp.739-745.

[13] Appala Srinuvasu Muttipati, Sangeeta Viswanadham, Radhika Senapathi , K.V.Brahmaji Rao. "Recognizing Credit Card Fraud Using Machine Learning Methods," *Turkish Journal of Computer and Mathematics Education*. 2021; vol.12, pp. 5-7.

[14] Talha Mahboob Alam, Kamran Shaukat, Ibrahim A. Hameed, (Senior Member, IEEE), Suhuai Luo, Muhammad Umer Sarwar, Shakir Shabbir, Jiaming , and Matloob Khushi. "An Investigation of Credit Card Default Prediction in the Imbalanced Datasets," *IEEE Access.* 2020; pp. 1-6.

[15] Ahmed Saad Hussein, Tianrui Li, Chubato Wondaferaw Yohannese, Kamal Bashir. "A New Preprocessing Approach for Highly Imbalanced Datasets by Improving SMOTE," *International Journal of Computational Intelligence Systems.* 2019; Vol.12, pp. 1412 - 1422.

[16] Ke G, Meng Q, Finley T, et al. "Lightgbm: A highly efficient gradient boosting decision tree[C]," *Advances in neural information processing systems.* 2017; pp. 3146-3154.
[17] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer. "SMOTE: Synthetic Minority Over-sampling Technique," *Artificial Intelligence Research*. 2002; vol. 16, pp. 321–357.