A new approach for credit card fraud detection using Machine Learning

Mitul Biswas[1], Swapan Debbarma[2]

{mitulbiswaschottu@gmail.com[1], swapanxavier@gmail.com[2]}

NIT Agartala, Tripura[1,2]

Abstract. The financial industry is growing at a rapid pace, and as a result, banking online transactions are on the rise as the government promotes digital transactions. Debit or credit cards have been used for the majority of financial transactions. As a result, the fraud associated with it is also on the rise. However, our current machine learning approach is unable to correctly detect fraudulent transactions since present fraud detection machine learning algorithms are taught and then evaluated on extremely unbalanced data sets, reducing their performance in real-world circumstances. In this paper, we proposed an algorithm that may work better after converting these imbalanced data sets into balanced data sets by using the oversampling technique so that system is not biased when the algorithm is actually implemented. The results show that the Random Forest algorithm with SMOTE performs better than the Logistic Regression machine learning algorithm.

Keywords: Machine learning, Fraud detection, Credit card fraud, SMOTE.

1 Introduction:

Information safety is already becoming increasingly critical as we get closer to a digital world. Detecting anomalous activity is the hardest challenge when it comes to information protection. When we are doing any kind of online transaction, a large number of people prefer to use credit cards. Even if we don't have the finances accessible right now, payment card credit limits might occasionally help us make purchases. Cyber attackers, on the other hand, make use of these qualities. To address this issue, we need a system that can stop a transaction if it detects something suspicious. This necessitates the development of a system that can track the pattern of all transactions and, if any pattern is aberrant, terminate the transaction. We now have a plethora of machine learning techniques that can assist us in classifying unusual transactions. The only requirements are historical data and converting those into balanced data and an appropriate algorithm that can better match our balanced data.

1.1. Classification of credit card fraud:

Application fraud: It happens when a fraudster gets control over any transactional application. Card not present: This kind of fraud happens when there is no physical card involved in the transaction. Lost or stolen card: When a card is lost then any people can misuse the lost or stolen card. Account theft: If a fraudster steals the account information and misuses the card. Fake website: If we have done a payment without knowing the originality of the website. The website may steal the card's information and misuse it. Merchant problem: In this fraud happens when

THEETAS 2022, April 16-17, Jabalpur, India

DOI 10.4108/eai.16-4-2022.2318073

a merchant misuse and share the information of the card with a third-party institution. Credit card fraud mainly happens when a card holder's details or card's details are leaked with any other fraudulent group without taking permission from the card owner, some transactions happen then we can say fraud happens. The financial sector is increasing and developing day by day and as a result banking online transactions are also in- creasing because govt. is also promoting digital transactions as Digital India. Most of the banking transactions happen through debit cards or credit cards. So, as a result, the fraud related to it also growing day by day. But our existing machine learning technique cannot able to accurately detect fraudulent transactions as our existing fraud detection machine learning techniques are trained on highly imbalanced data sets. So, the accuracy of these techniques in the real-life scenario is varied highly. So, my goal is to convert these imbalanced data set into balanced data set and find out which of the existing algorithms or techniques give the best results in all the fraudulent scenarios with the help of pattern recognition and machine learning.

2 Problem statement:

Previous authors have used machine learning algorithms to work on imbalanced datasets of credit card frauds. No authors have converted the imbalanced datasets to balanced datasets and used the oversampled or undersampled data for the training and testing phase of machine learning algorithms to find out which machine learning algorithm gives us the best results in real case scenarios to find fraudulent transactions.

3 Related works:

Asha RB [1] uses support vector machine (SVM), k-nearest neighbour (KNN), and artificial neural network (ANN) methodologies to discover credit card fraud, as well as advanced analytics and deep learning models to find non-fraud transactions.

Another paper by K. VENGATESAN and A. KUMAR [2] suggests credit card fraud detection in the banking industry using a machine learning algorithm and compares the accuracy of logistical regression and KNN techniques. The banking industry provides a variety of services to its consumers, including ATM cards, Internet banking, gold loans, education loans, card payments, and credit cards, in order to entice people to create bank accounts. customers often use credit cards 24 hours a day, so the bank server may utilize machine learning algorithms to keep track of all transactions. It should be able to locate or anticipate fraud detections. We need to categorize if each transaction is lawful or not using the data set, which contains all of the attributes of each transaction.

A credit card fraud detection model was created by Alejandro Correa Bahnsen, Djamila Aouada, Aleksandar Stojanovic, and Bjorn Ottersten [3], and extracting the proper features from transactional data is crucial. In order to uncover customer purchasing trends, this is frequently achieved by aggregating transactions. We propose to develop a new set of attributes based on the von Mises distribution, which they used to examine the periodic behavior of transaction time in their research.

Krishna Modi [4] conducted a comparison study of numerous strategies for detecting fraud, including decision trees, rule-based mining, artificial neural, fuzzy-based clustering approaches, hidden Markov models, and hybrid approaches of these methods.

M. Suresh Kumar, V. Soundarya, S. Kavitha, E.S. Keerthika, E. Aswini [5] have developed a system for detecting fraudulent transactions and determining their correctness, the suggested system employs the Random Forest Algorithm (RFA). This technique utilizes decision trees to classify the dataset and is based on supervised learning. A confusion matrix is produced once the dataset has been classified. The confusion matrix is used to judge the performance of the Random Forest Algorithm.

John O. Awoyemi, Adebayo O. Adewunmi, Samuel A. Oluwadare [6] have done a comparative study where the performance of nave bayes, k-nearest neighbor, and logistic regression on highly skewed credit card fraud data is investigated.

Dejan Varmedja, Mirjana Karanovic, Srdjan Sladojevic, Marko Arsenovic, Andras Anderla [7] have done research that confirms several methods for identifying transactions as fraudulent or legitimate. Their proposed model may be used to detect additional anomalies.

Alejandro Correa Bahnsen, Djamila Aouada, Aleksandar Stojanovic, Bjo¨rn Ottersten [8] extend the transaction aggregation method by proposing a whole fresh batch of characteristics based on evaluating the periodic behavior of transaction time using the von Mises distribution.

Eesha Goel, Abhilasha, Ankit Agarwal [9] have made a system using the Random Forest algorithm (RFA) to do the analysis of fraud in online shopping, they detect the frauds and prevent unusual activities.

S. Monika, K. Venkataramanamma, P. Pritto Paul, M. Usha [10] used a Random Forest algorithm for fraud detection.

Devi Meenakshi., Janani., Gayathri., Mrs. Indira. N [11] developed a technique in which the accuracy of detecting fraud might be increased by employing a random forest algorithm. The random forest algorithm's classification procedure is used to analyze the data set and the user's current dataset. Finally, improve the precision of the output data. The approaches' accuracy, sensitivity, specificity, and precision are used to evaluate their performance.

Sudeep Dogga [12] used the Random Forest algorithm (RFA) to detect credit card fraud detection whether the transaction is genuine or not. It is capable to solve both classification and regression issues.

4 Existing system:

Machine learning algorithms like Random Forest, K-Nearest neighbor, SVM, Logistic Regression methods have been used to differentiate between fraudulent and genuine transactions of credit cards. But all the authors have taken highly imbalanced data sets of credit card frauds in the training and testing phase. So, the accuracy of all the existing machine learning algorithms is varied based on different authors' experiments.

5 Proposed system:

Machine learning algorithm performance depends on the data set provided to the algorithm during the learning time of the algorithm. Based on the given test data the algorithm learned itself and make a decision based on this learning whenever the algorithm get the input during real-time application. As we all know in the current time if we consider the Business Standard report [14] a total number of banking transactions are 41M and a total number of frauds may report as per the report of New India Express [13] is 1,194. So, if we consider the ratio between fraudulent transactions over a total number of transactions. So, this is a highly imbalanced data set if we are considering data set formed by real-time transactions. Because of the significantly skewed class distribution, imbalanced classification is particularly difficult as a predictive modeling assignment. This is why standard machine learning algorithms and assessment measures that presume a balanced class distribution perform poorly. This is a great challenge for machine learning algorithms to predict frauds properly. Although there are few proposals of different algorithms which may work on imbalanced data sets with high accuracy. One-class SVM (Support Vector Machine), for example, is a classification technique that aids in the detection of outliers in data. This method may be used to deal with challenges with unbalanced data, such as fake detection [15]. Random Forest algorithm is another approach that may deal with unbalanced data sets [16]. So, we need to convert the availableimbalanced data set into a balanced data set as far we know without a properly balanced data set, we cannot accurately measure any kind of fraudulent transactions as we cannot properly train the machine learning algorithm.

6 Methodology:

This section is intended to discuss implementation, which covers the methods and other components required for the proposed system's implementation. The implementation in this article will begin with importing the dataset and then cleaning and normalizing it. Dataset is divided into two portions to use for the training phase and testing phase. Finally, the system may be able to detect fraudulent and nonfraudulent transactions. We must employ Python as a computer language in the suggested system. Python is a scripting language that is simple to use, interpreted, object-oriented, and high level. Python is a powerful language of programming for building machine learning algorithms. It offers a number of important modules and standard functions for machine learning.

Numby: It's a member of the Python libraries. The most common jobs are multidimensional groups and algebraic equation calculations. Pandas: A standard library, such as Pandas, is an example. Pandas is mostly used for information collecting and modification. It's usually used to read and transfer the information from a dataset. Scikitlearn: It is a Python package that may be used for computational intelligence algorithms. It is the most appropriate Python module for computational analysis. Keras: Keras is a sophisticated application programming interface. It's a programming interface for neural networks. It is possible to execute it on top of a tensor flow. Computational intelligence algorithms are typically implemented using it. It can run on both the CPU and the GPU simultaneously. Keras and a backend running tensor circulation were

employed in this work. Keras with the tensor flow as a backend aid in neuromorphic structure training. Mysql: For storing, a MySQL database is utilized. For storing user information, we utilized MySQL. The user must enrol via submitting credentials, which are then saved in the database server. Tkinter: Tkinter is a user interface library written in Python. It can run on both Unix and Windows systems. We may create it by first installing the Tkinter module, then creating a GUI, and then adding several panels and invoking them in the cycle.

Firstly, the imbalanced datasets are transformed into a balanced dataset using SMOTE and then the machine learning algorithms are used to check which algorithm is giving the best precision, recall and accuracy, and f1 score. Previous authors have never tested the oversampled and undersampled data sets to run for the training and testing phase. The imbalanced datasets are highly biased towards the majority classes. That is why we have to make a dataset of 50/50 fraudulent and non-fraudulent datasets and then run them in the training and testing phase of the algorithms. SMOTE is used to generate new synthesized features in order to keep the categories balanced. This is used to resolve problems of unbalance. It is used mainly because in order to achieve an appropriate balance between the smaller and the larger groups, it produces synthesized credits from the smaller class. SMOTE also determines the distance between the smaller class's nearest neighbors and develops synthesized endpoints in between these ranges. It enables us to maintain more information since, unlike random sampling, we don't have to reject any rows. After all, SMOTE will be much more factual than random under-sampling because, as said before, no rows are completely removed during learning.

6.1. Mainly used algorithms for our experiments are:

Logistic regression.

Random Forest classifier.

K-Nearest neighbor.

Support vector machine.

The working processes of those above methods are described below:

6.2. Logistic Regression process:

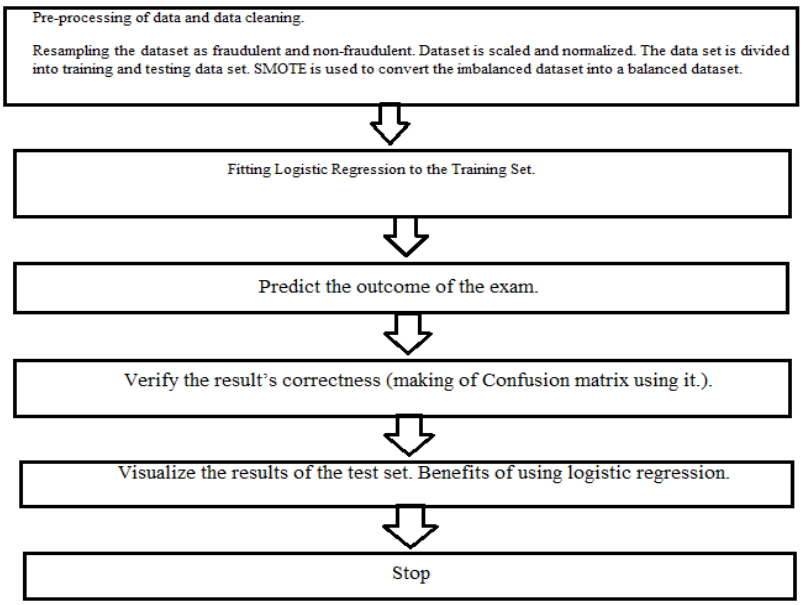

Fig.1. shows the working method of the Logistic Regression process.

6.3. Random Forest classifier process:

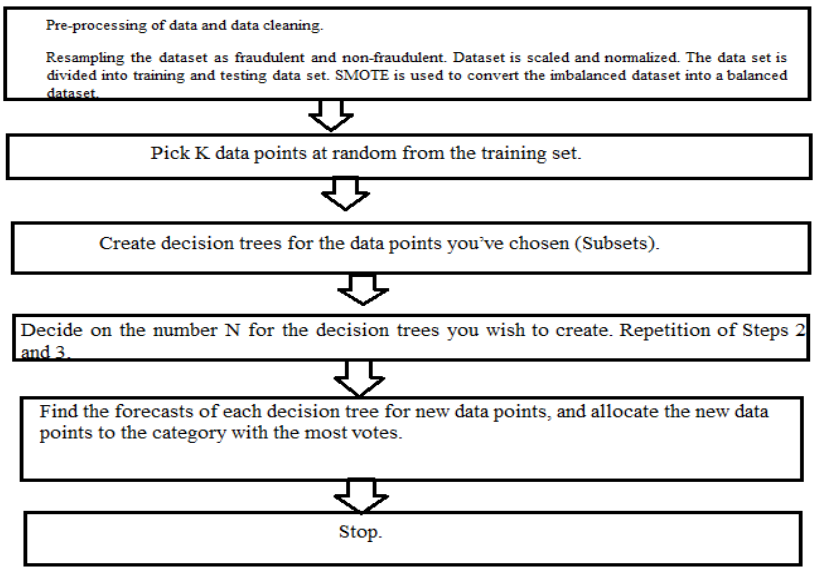

Fig.2. shows the working method of the Random Forest Classifier.

6.4. K-Nearest neighbor process:

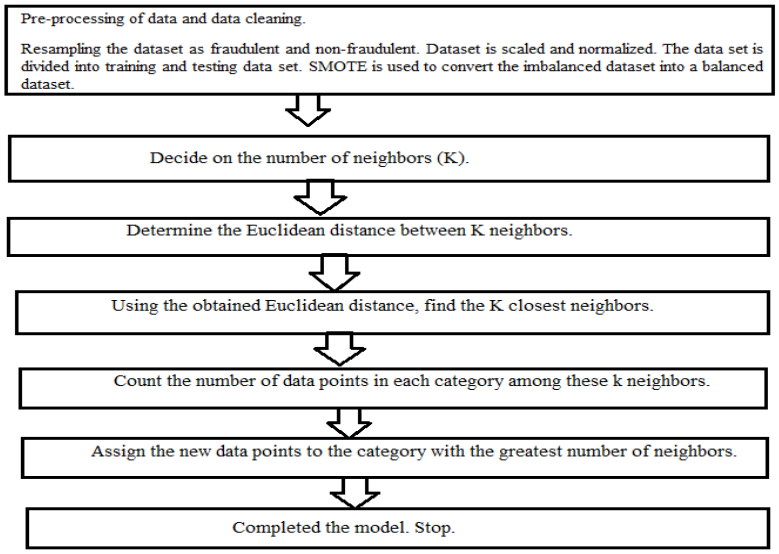

Fig.3. shows the working method of the K-Nearest Neighbor.

6.5. Support vector machine process:

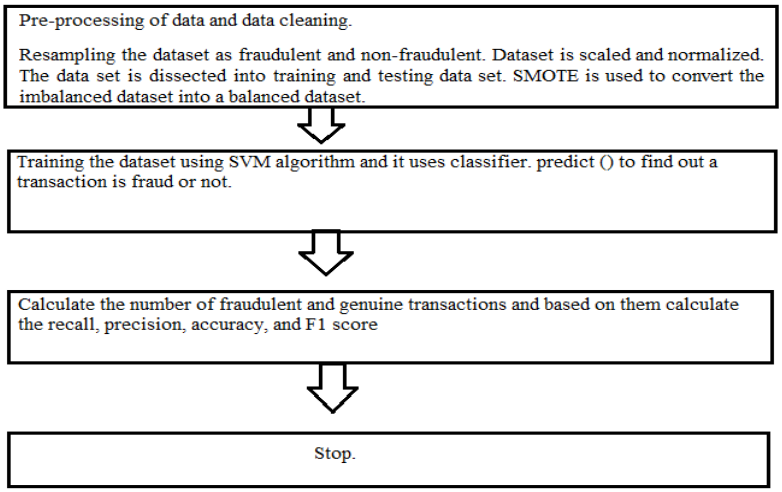

Fig.4. shows the working method of the Support vector machine (SVM).

7 Results and discussion:

The information used in the proposed mechanism may be acquired from the webpage www.kaggle.com. The operations done by clients of a European bank within 2013-14 are utilized as a sample. It is used to train and validate our system in order to detect fake and legit transactions.

Evaluation measure:

We will discuss the multiple ways to check the performance of the machine learning algorithms. We can check the efficiency of the algorithms based on the following parameters.

Confusion matrix:

The confusion matrix is basically the best representation of the following parameters.

True positives (TP)**:** Predicted positive and they are actually positive**.**

False positives (FP)**:** Predicted positive and they are actually negative.

True negatives (TN)**:** Predicted negative and they are actually negative.

False negatives (FN)**:** Predicted negative and they are actually positive.

Accuracy measurement = (TP + TN) / (TP + FP + TN + FN)

Precision measurement = TP / (TP + FP)

Recall measurement = TP / (TP + FN)

F1 Score measurement = (2 * Precision * Recall) / (Precision + Recall)

7.1. Results:

Table-1 shows the performance results of different ML methods:

Evaluation parameters in %	Logistic Regression	Random Forest	KNN	SVM
Precision	92.8956	95.9887	94.5891	93.228
Recall	93.112	95.1234	92.008	93.005
Accuracy	90.448	94.9991	94.999	93.963
F1 Score	92.112	95.1102	91.003	93.479

Table-1 [17] is taken as a reference of showing different performance measures of different machine learning algorithms which is giving results based on the imbalanced training and testing datasets. F1 score is taken mainly as it is the harmonic mean between recall and precision, it tells us how robust, and precise our classifier is. we are taking the F1 Score to measure the performance. Based on the above table mentioned F1 score, we can say that the Random Forest algorithm gives us better performance than the other three machine learning algorithms to find the fraud transactions using an imbalanced dataset.

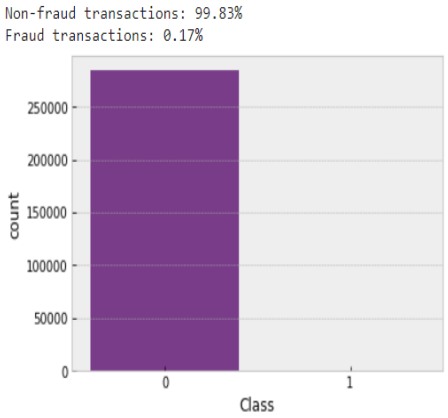

Fig.5.This shows that the taken dataset is consist of fraud and non-Fraud transactions**.**

Fig.5. shows us how imbalanced the dataset is. Dataset is made with 99.83% of genuine transactions and 0.17% of fraud transactions. So, we have to balance the dataset using oversampling, then the oversampled dataset will be used for the training and testing phase.

We will use transaction amount and transaction time distribution to understand how skewed these features in the dataset are. For the privacy policies, we can't get the names and other features. All other features went through PCA transformation which means that they are previously scaled. We will use the feature scaling method to scale the transaction amount and transaction time.

```
f, (ax1, ax2) = plt.subplots(1, 2, figsize=(13, 6))
ax1 = sns.distplot(df['Time'], ax=ax1, color='g')
ax2 = sns.distplot(df['Amount'], ax=ax2, color='y')
ax1.set_title('Distribution of Time', fontsize=13)
ax2.set_title('Distribution of Amount', fontsize=13)
```

```
Text(0.5, 1.0, 'Distribution of Amount')
```

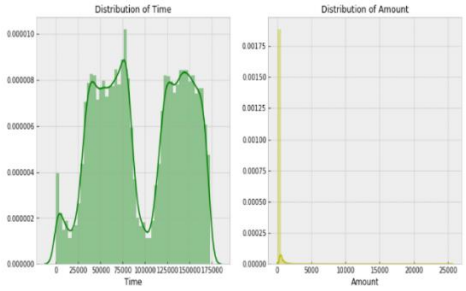

Fig.6. It shows the distribution of transaction time and amount

```
Transaction Number x_train dataset:  (227845, 30)
Transaction Number y_train dataset:  (227845, 1)
Transaction Number x_test dataset:  (56962, 30)
Transaction Number y_test dataset:  (56962, 1)
Before OverSampling, counts of label '1': [391]
Before OverSampling, counts of label '0': [227454]

After OverSampling, the shape of train_x: (454908, 30)
After OverSampling, the shape of train_y: (454908,)

After OverSampling, counts of label '1', %: 50.0
After OverSampling, counts of label '0', %: 50.0
]: <matplotlib.axes._subplots.AxesSubplot at 0x7fc6aaaf4b38>
```

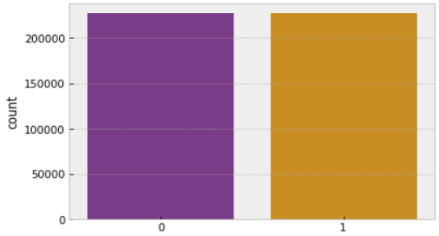

Fig.7. It shows that the transactions dataset of fraud and non-fraud transactions is now balanced after oversampling.

Fig.7. shows us the results after the imbalanced dataset is balanced. The results now show that the dataset now consists of 50% of original transactions and 50% of fake transactions. SMOTE is invoked after cross-validation is done, so that overfitting and data leakage problems will not happen. Now, this dataset will be used to train and test the machine learning algorithms.

7.2. Logistic Regression with SMOTE:

	precision	recall	f1-score	support
0	1.00	0.98	0.99	56861
1	0.06	0.94	0.12	101
accuracy			0.98	56962
macro avg	0.53	0.96	0.55	56962
weighted avg	1.00	0.98	0.99	56962

Fig.8. It shows the results of different measurement matrices after Logistic Regression is run with SMOTE.

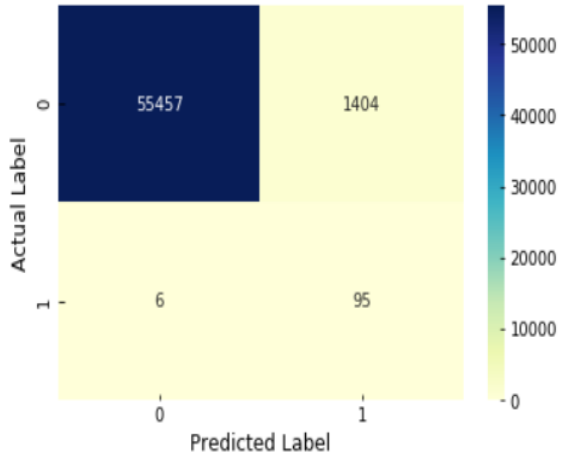

Fig.9.It shows the confusion matrix of Logistic Regression with SMOTE results.

```
y_pred_prob = logreg.predict_proba(x_test)[:,1]
precision, recall, thresholds = precision_recall_curve(y_test, y_pred_prob)
plt.plot(precision, recall)
plt.xlabel('Recall')
plt.ylabel('Precision')
plt.title('Precision Recall Curve')
```

```
Text(0.5, 1.0, 'Precision Recall Curve')
```

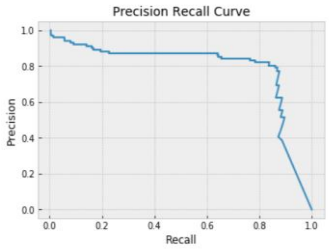

Fig.10. It shows the Precision-Recall curve of Logistic Regression with SMOTE.

It shows a high recall value that means the model is able to find out the highest number of fraud transactions, the precision shows very low which is not good because it means the model classifies many genuine transactions as fraud transactions. It is very important to have a good precision value.

7.3. Random Forest with SMOTE:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	56861
1	0.88	0.83	0.86	101
accuracy			1.00	56962
macro avg	0.94	0.92	0.93	56962
weighted avg	1.00	1.00	1.00	56962

Fig.11. It shows the results of different measurement matrices after Random Forest is run with SMOTE.

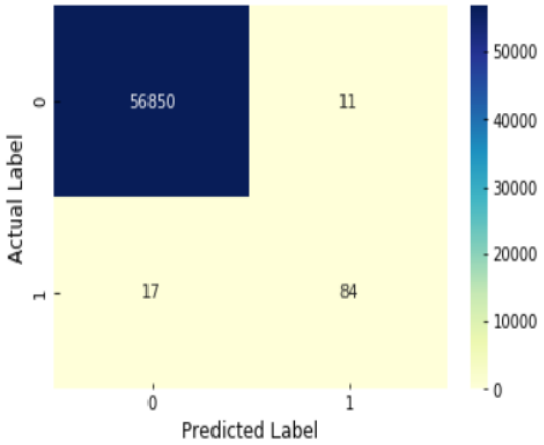

Fig.12. It shows the confusion matrix of Random Forest with SMOTE results.

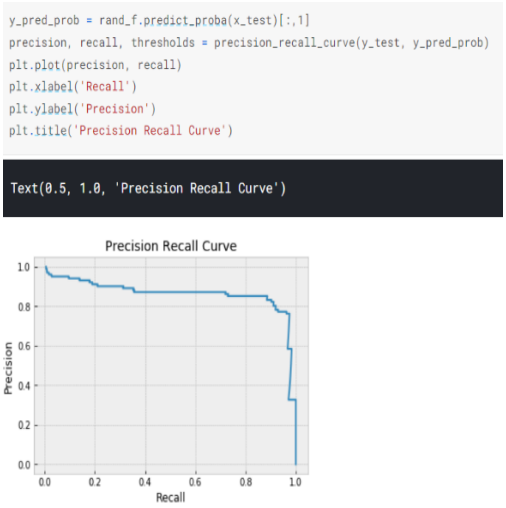

Fig.13. It shows the Precision-Recall curve of Random Forest with SMOTE.

Random Forest with SMOTE shows us better performance than Logistic Regression with SMOTE. It gives us a better and high recall and precision value. The recall is gone lower a little bit but also the precision is increased significantly which means a lot in the case of fraud detection and as we knew that it's a trade-off.

8 Conclusion:

Good prediction results can be achieved by using balanced datasets. In this paper, the Random Forest classifier gives us the best results with balanced datasets. It helps us to detect more than 80% of fraud transaction cases and at the same time, it is not classifying many non-fraud transaction cases as fraud. In this paper, SMOTE is used with Random Forest classifier and Logistic Regression only and gives us the results that Random Forest with SMOTE gives us the best results. Random Forest shows the best results both with balanced and imbalanced datasets. There is always a trade-off between recall and precision, it always depends upon situation objectives to decide which approach is the best in each particular situation.

9 References:

[1] RB. Asha, S. Kumar Credit card fraud detection using artificial Neural network. In Elsevier,2666-285X/2021, Publishing services by Elsevier B.V., Global transactional proceeding 2 35-41, https://doi.org/10.1016/j.gltp.2021.01.006, 23rd January 2021.

[2] K. VENGATESAN, A. KUMAR, S. YUVRAJ, V. D. AMBETH KUMAR, AND S. S. SABNIS Advances in Mathematics: Scientific Journal 9 (2020), no.3, 1185–1196 ISSN: 1857-8365 (printed); 1857-8438 (electronic) https://doi.org/10.37418/amsj.9.3.43 Spec. Issue on RDESTM, 2020.

[3] Alejandro Correa Bahnsen, Djamila Aouada, Aleksandar Stojanovic, and Bjorn Ottersten Detecting Credit Card Fraud using Periodic Features. In IEEE 14th International Conference on Machine Learning and Applications, 978-1-5090-0287-0/15, 2015.

[4] Krishna Modi, Reshma Dayma Review on Fraud Detection Methods in Credit Card Transactions. In International Conference on Intelligent Computing and Control (I2C2'17), 2017.

[5] M. Suresh Kumar, V. Soundarya, S. Kavitha, E.S. Keerthika, E. Aswini CARD FRAUD DETECTION USING RANDOM FOREST ALGORITHM. IEEE 3rd International Conference on Computing and Communication Technologies ICCCT 2019, 978-1-5386-9371- 1/19, 2019.

[6] John O. Awoyemi, Adebayo O. Adewunmi, Samuel A. Oluwadare Credit card fraud detection using Machine Learning Techniques. In IEEE Conference 978-1-5090-4642- 3/17, 2017.

[7] Dejan Varmedja, Mirjana Karanovic, Srdjan Sladojevic, Marko Arsenovic, Andras Anderla Credit Card Fraud Detection - Machine Learning methods. In IEEE Conference 978-1-5386-7073-6/19/, 2019.

[8] Alejandro Correa Bahnsen, Djamila Aouada, Aleksandar Stojanovic, Bjo¨rn Ottersten, Feature engineering strategies for credit card fraud detection. In Correa Bahnsen et al. / Expert Systems with Applications 51 (2016) 134–142, http://dx.doi.org/10.1016/j.eswa.2015.12.030 0957-4174/Elsevier Ltd., 2016.

[9] Eesha Goel, Abhilasha, Ankit Agarwal Fraud Detection Using Random Forest Algorithm. In Eesha Goel et al. / International Journal of Computer Science Engineering (IJCSE), Vol. 5 No.05, Sept 2016.

[10] S. Monika, K. Venkataramanamma, P. Pritto Paul, M. Usha Credit card fraud detection using random forest algorithm. In International Journal of Research in Engineering, Science and Management Volume-2, Issue-3, March 2019.

[11] Devi Meenakshi., Janani., Gayathri., Mrs. Indira. N CREDIT CARD FRAUD DETEC- TION USING RANDOM FOREST. In International Research Journal of Engineering and Technology (IRJET), Volume: 06 Issue: 03, March 2019.

[12] Sudeep Dogga The Role of Random Forest in credit card fraud analysis. In Bournemouth university England.

[13] The New India Express NCRB report 2020: Debit credit card fraud climbs steeply, more cases compared to 2019. In https://www.edexlive.com/news/2021/sep/16/ncrb-report- 2020-debit-credit-card-fraud-online-climbs-steeply-225-more-cases-when-compared-to- 2-24053.html, 2020.

[14] Business Standard India leads the world approximately 41M transactions in a day. In https://www.edexlive.com/news/2021/sep/16/ncrb-report-2020-debit-credit-card-fraud- online-climbs-steeply-225-more-cases-when-compared-to-2-24053.html, 2020.

[15] Roman Chuprina, Olena Kovalenko Credit card fraud detection: Top ML solutions 2021. In Card Fraud Detection: Top ML Solutions in 2021 (spd.group)., 2021.

[16] Lukas Frei Detecting credit card fraud using machine learning. In Detecting Credit Card Fraud Using Machine Learning — by Lukas Frei — Towards Data Science.

[17] Naresh Kumar Trivedi, Sarita Simaiya, Umesh Kumar Lilhore, Sanjeev Kumar Sharma An Efficient Credit Card Fraud Detection Model Based on Machine Learning Methods, In International Journal of Advanced Science and Technology Vol. 29, No. 5, pp. 3414 – 3424, 2020.