e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:12/December-2022 Impact Factor- 6.752 www.irjmets.com

HYBRID APPROACH FOR OPTIMIZED FRAUDULENT CREDIT CARD TRANSACTION DETECTION USING MACHINE LEARNING

Sai Deep I[*1], Alapati Naga Praneeth[*2], Kowshik PJ[*3], Bellamkonda Surya Kiran[*4], Narasimhayya BE[*5]

[*1,2,3,4]B.Tech CSE - AI, Department Of Computer Science And Technology Faculty Of Engineering And Technology JAIN UNIVERSITY, Karnataka, India.

[*5]Associate Professor, Department Of CSE-AI, Faculty Of Engineering And Technology JAIN UNIVERSITY, Karnataka, India.

DOI : https://www.doi.org/10.56726/IRJMETS31968

ABSTRACT

The target of our study is obtaining a method for intruder detecting system that offers an alternative safety for the protection of the most commonly used financial transaction model. There is a large range of creditcard transactions occurring within the actual world but there is additionally the 0.33% person who is monitoring our activities and puts people in problems. This changed into going on not best now however additionally approximately 20 years in the past, and with the help of the brand-new tech algorithms, this fraud hobby has been multiplied hastily in certain areas consisting of electronic shopping, advertising and marketing, and so forth. So, to locate those styles of fraud pastime, our research work is figuring out the share of the fraudulent inside the given statistics set. in recent times, technology is improvised and frauds had been growing unexpectedly inside the banking zone, fraudulent sports in credit score-card were improved. Our research paints, the process is to make correct predictions while balanced statistics is fed. With using the gadget studying [ML], this research examine analyses and summarizes the frauds in the credit card transactions. two library documents is inclusive of PyCaret and SMOTE had been used for records balancing five getting to know algorithms accomplished the detection of fraudulent hobby, among them random wooded area algorithm which uses a balanced dataset of SMOTE offers the best estimate of the generalization mistakes and to be resistive to overfitting.

Keywords: SMOTE, Pycaret, Imbalanced Data, Precision, Recall, F1 Score, XGBoost.

I. INTRODUCTION

Over the past decade, e-commrce has been increased and credit card usage has increased significantly. The increased use of credit cards has led to a steady increase in the fraudulnt transactions. Fraudulent transactions have had a serious impact on the financial industry. According to a report, about $27.85 billion was lost to credit card frauds in 2018, up 16.2% from $23.97 billion in 2017, to $35 billion by 2023. estimated to reach. These kinds of losses can be reduced through effective fraud monitoring & prevention.

Machine learning, on the other hand, has been used to develop several systems for detecting credit card fraud. However, creditcard fraud detection remains a learning challenge due to class imbalance in the dataset. Class imbalance is not the only problem preventing credit card fraud from being detected, but it is the most significant. This is a problem which arises in real-world ML applications where the dataset has an uneven class distribution.

Most of the credit card transaction records are imbalanced, with the number of legitimate transactions far outweighing the number of fraud transactions. The most traditional ML algorithms perform well on balanced data class distributions skew the performance of traditional ML algorithms against the majority class, because they are designed to consider the error rates rather than the class distributions. Therefore, more examples of the minority class are misclassified than examples of the majority class. These motives leading to fraud are criminal purposes. This practice of pursuing fraud is essentially aimed at siphoning money illegally, resulting in loss of financial or personal gain. Credit card transactions can be organized physically or virtually. In physical transactions, users exchange cards personally at the time of purchase. Virtual Transactions include online operations via the World Wide Web.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:12/December-2022 Impact Factor- 6.752 www.irjmets.com

II. METHODOLOGY

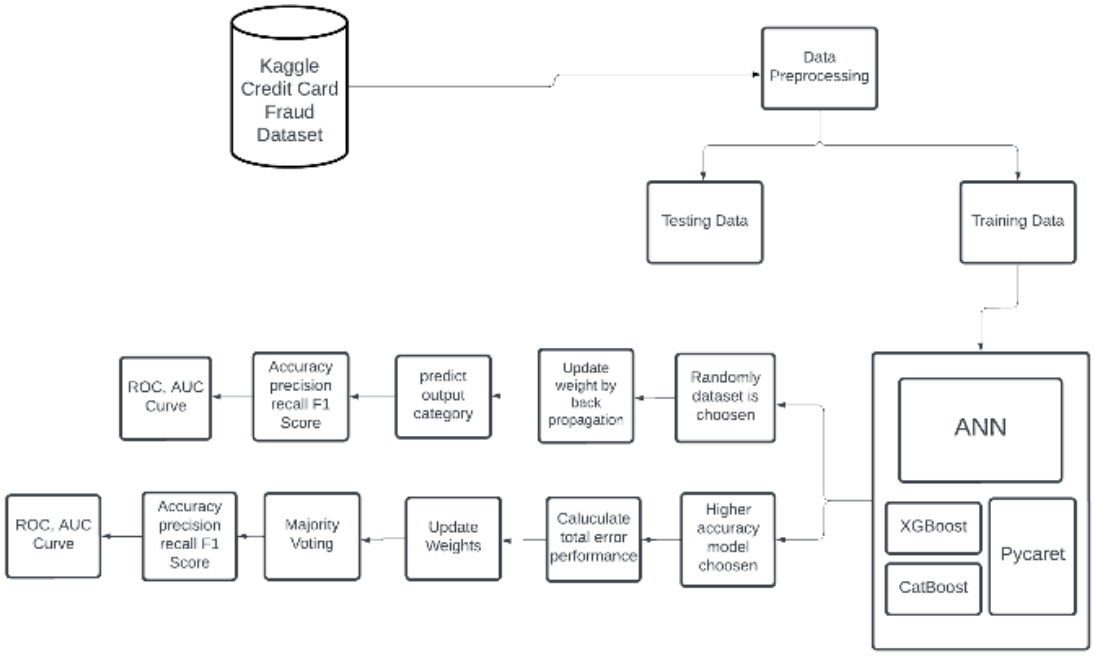

Figure 1: Project Architecture

Figure 1 demonstrates the workflow implemented in python using machine learning and deep learning. The dataset is trained using Artificial neural network and algorithms from Machine Learning classifiers like logistic regression, Random forest. Pycaret library is used to analyze the performance of all types of classifiers and displays the evaluation metrics. From pycaret it is observed that the models like Xgboost, Catboost and random forest classifier has the highest precision and recall scores compared to other models.

2.1 Artificial neural networks:

This algorithm makes use of an intricate connection between the output data and the input data to reveal the hidden correlated patterns. We make use of this strategy for recognizing and predicting the un-authorized transactions. The model's innovative capabilities can increase the existing data analysis techniques. As displayed in the figure the layers which are at the initial state are the input layers which receives the information from the data set which can be any kind of data format like text, audio, video, image and other types. The network entails of dense layers which are hidden layers made of units that transmutes the given input data into a desired output upon various mathematical computations by recognizing the patterns. There exist some random weights(w) which are assigned to every neuron initially. The weights are multiplied by the provided inputs(x) attained from the input layer. The weighted sum is formed by aggregating the multiplied values. The stimulating task converts it into a subsequent output. Once the network receives the input variable quantity, the weight of the input is allocated at random. The significance of every input data point is implied by its weight in terms of estimating the outcome. The discrimination parameter, on the other side, provides an option to update the activation curve. Once all the products are aggregated, we obtain the weighted sum. In this scenario we feed the input perceptron with the existing data of fraud transactions which is undergone through Principal Component Analysis. This input data is multiplied with random weights(W) initially to form an aggregated sum. These random weights are updated during the process of back propagation. Each perceptron is triggered with an activation function once the aggregated sum is calculated. In the output layer the probability is determined which classifies the category of the given input.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:12/December-2022 Impact Factor- 6.752 www.irjmets.com

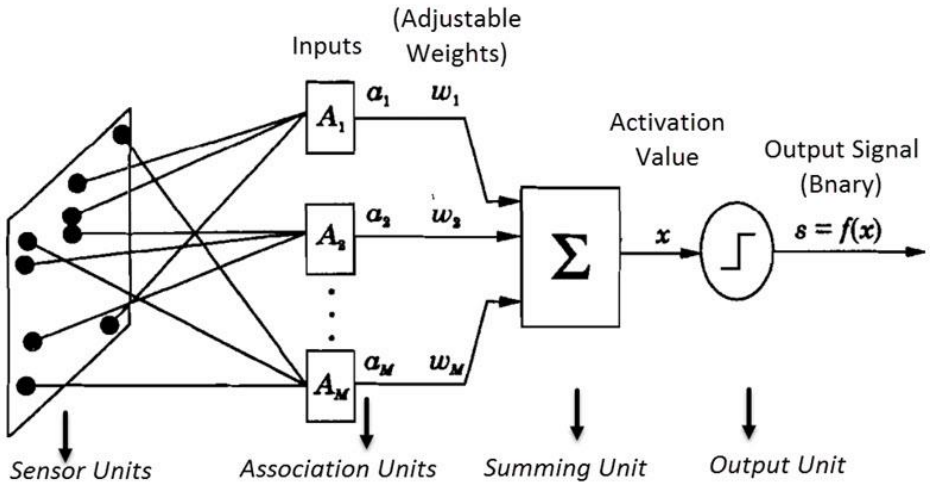

Figure 2: ANN Architecture

2.2 Pycaret:

A simple, easy to implement open-source library that automates all the required classification algorithms. It is a low-code machine learning library that aims to decreases the cycle time from hypothesis to intuitions. It enables to achieve iterative end-to-end data science researches. It is a wrapper around several machine learning libraries and frameworks such as scikit-learn, XGBoost. All of the procedures performed in typical experiment are implemented and necessarily stored in a custom pipeline.

Pycaret is initialized and imported with all modules required. In setup we use setup() function configure simple data preparation like scaling , power transforms. It will expect if all the variables are as per the dataset, as per requirements. Give input as 'y' and enter. Input for percentage of data that should be used for training will be asked for the training model. Here we have the split data as 70/30.

Finally we can compare the trained models using compare function which gives us the report of all models summary of their performance.

After comparing all the models and it's metrics we choose the best-performing model and its configuration.

2.2.1 CatBoost Classifier:

CatBoost is one of the recently released machine learning algorithm. It integrates with deep learning frameworks and can process different types of data to solve a wide variety of business problems. CatBoost works well with multiple categories of data such as text, images, and audio. Boost comes from the gradient boosting machine learning algorithm. Gradient boosting is a unique and efficient algorithm widely used for many different kinds of problems and challenges. It has the ability to deliver superior results compared to deep learning models, which require large amounts of data for accuracy. CatBoost converts categorical class variables to integers using statistical combinations of categorical and numeric features, so it can be used without pre-processing techniques. There are several configurable parameters such as: B. Learning rate, tree depth, folding size, sagging temperature.

2.2.2 XGBoost Classifier:

An optimized distributed boosting library which is designed to be highly efficient and flexible. It is an ensemble learning algorithm as it combines the results of multiple models. The decision trees are created in a sequential form where weights have a major importance. The weights are initialized to the independent variables which are given as input to the decision tree. The weights which are predicted erroneous is modified and are the fed to the second decision tree. These unique classifiers ensemble to form a robust precise model. It can work for multiple purposes like regression, classification, ranking and custom-defined challenges.

The scores for the prediction of every distinct decision tree sum up to get the final score.

Mathematically the model can be represented in the form:

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:12/December-2022 Impact Factor- 6.752 www.irjmets.com

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

Where K is the count of trees, f is functional space of F.

The objective of the XGBoost model is given by the following equation:

$$obj(\theta) = \sum_i^n l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Which is to minimize the magnitude of loss function by applying additive strategy.

SMOTE: SMOTE is an oversampling technique which generates samples for the minority class. Using this algorithm, we overcame the overfitting problem posed by random oversampling. SMOTE uses KNN to generate synthetic samples. We tweaked the model increasing false positives and reduces the false negatives**.** By using SMOTE, we can increase recall but it effects precision negatively.

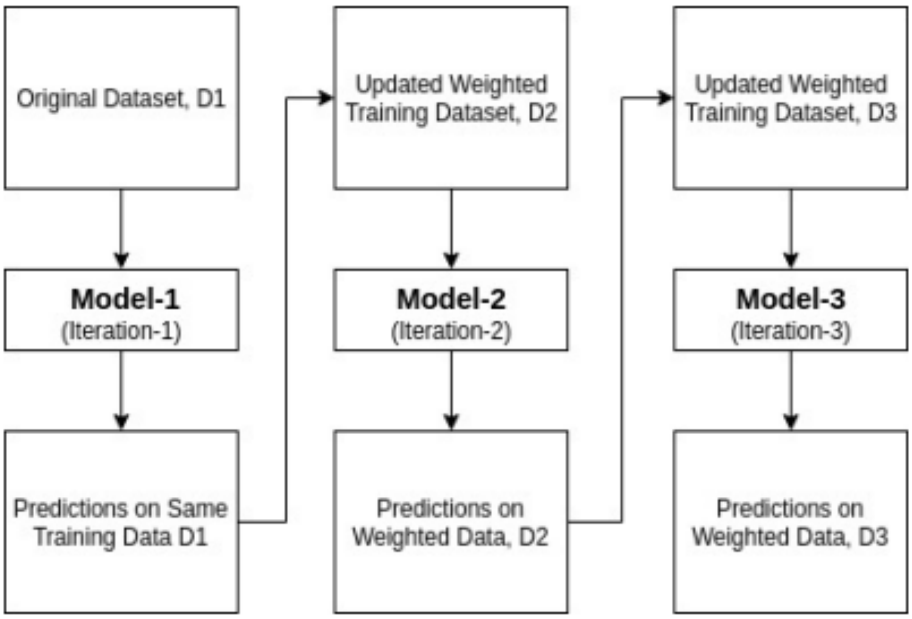

Figure 3: XGBoost Architecture

III. MODELING AND ANALYSIS

In the dataset used for testing has a total of 56844 transactions. By using this dataset we can see our performance of model in real time transactions.

Following is the modelling of neural network

```
#0.1 test size
y_pred = (model_ann.predict(X_test) > 0.5).astype("int32")
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

```
1781/1781 [==============================] - 2s 1ms/step
[[56844    20]
 [   18    80]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56864
           1       0.80      0.82      0.81        98

    accuracy                           1.00     56962
   macro avg       0.90      0.91      0.90     56962
weighted avg       1.00      1.00      1.00     56962
```

Figure 4: Metrics of ANN

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:12/December-2022 Impact Factor- 6.752 www.irjmets.com

```
loss_df = pd.DataFrame(model_ann.history.history)
loss_df.head()
```

	loss	accuracy	val_loss	val_accuracy
0	0.009699	0.999249	0.002919	0.999605
1	0.003693	0.999361	0.003627	0.999473
2	0.003568	0.999356	0.003254	0.999342
3	0.003476	0.999371	0.002511	0.999605
4	0.003468	0.999395	0.002548	0.999561

Figure 5: Summary of ANN metrics

From the above result of Neural network model. We can see that we got highest accuracy and precision of 99.96%.

Following is the confusion matrix after comparing the predicted data with the test data.

The biggest disadvantage when applying it is the problem with an imbalanced dataset because the number of fraud transactions in real life is much less than non-fraud cases. This problem is partially solved by the data resampling technique. For future works could be used improving classification algorithm to fit better with imbalanced dataset.

	0	1
0	56844	20
1	18	80

IV. RESULTS AND DISCUSSION

```
# we will get Accuracy, Recall, Preccision, F1-score..... for different models.

compare_models()
```

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa
0	CatBoost Classifier	0.9996	0.973	0.801	0.9475	0.866	0.8658
1	Random Forest Classifier	0.9995	0.9271	0.7552	0.9406	0.835	0.8348
2	Extra Trees Classifier	0.9995	0.9547	0.78	0.9458	0.8526	0.8524
3	Extreme Gradient Boosting	0.9995	0.9749	0.8008	0.9181	0.8535	0.8532
4	Linear Discriminant Analysis	0.9994	0.9822	0.7675	0.8787	0.816	0.8158
5	Ada Boost Classifier	0.9992	0.9635	0.7093	0.8195	0.7551	0.7547
6	Logistic Regression	0.9991	0.9116	0.7052	0.7555	0.7266	0.7262
7	Decision Tree Classifier	0.9991	0.892	0.7845	0.7334	0.7561	0.7557
8	Gradient Boosting Classifier	0.999	0.7003	0.5433	0.8143	0.6402	0.6397
9	Ridge Classifier	0.9988	0	0.4067	0.8377	0.5428	0.5423
10	K Neighbors Classifier	0.9983	0.5708	0.025	0.45	0.0463	0.0462
11	SVM - Linear Kernel	0.998	0	0	0	0	-0.0005
12	Light Gradient Boosting Machine	0.996	0.5455	0.2792	0.1761	0.1913	0.1897
13	Naive Bayes	0.9924	0.9704	0.6183	0.1332	0.2189	0.2167
14	Quadratic Discriminant Analysis	0.9752	0.9661	0.8798	0.0584	0.1095	0.1066

Figure 6: Accuracy Summary

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:12/December-2022 Impact Factor- 6.752 www.irjmets.com

Following are the list of different machine learning models that have been run here using PyCaret and it is observed that CatBoost Classifier gets the best model with an accuracy of 0.9996, an accuracy of 0.94, an F1 score of 0.86, and a kappa score of 0.86.

We used 70% of the data to train the model, and the remaining 30% is used to test the model and predict its outcome. Here predict model is used to query the result for both catboost and XG boost.

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa
CatBoost Classifier	0.9996	0.973	0.801	0.9475	0.866	0.8658
Extreme Gradient Boosting	0.9995	0.9749	0.8008	0.9181	0.8535	0.8532

Figure 7: Concluding Models Accuracy

V. CONCLUSION

In this project, it was found that we can determine the good performing model using Pycaret. It was found that CatBoost and Extreme Gradient (XG Boost) are the most effective compared to other methods. Hence we used these algorithms for detection of fraud transactions. This research paper gives a detailed explanation of our project and our approach.

ACKNOWLEDGEMENTS

The authors are grateful for the guidance of Prof. Narasimhayya B E throughout the implementation of the project.

VI. REFERENCES

[1] Fayyomi, Aisha & Eleyan, Derar & Eleyan, Amina. (2021). A Survey Paper On Credit Card Fraud Detection Techniques. International Journal of Scientific & Technology Research. 10. 72-79.

[2] Asha RB, Suresh Kumar KR. Credit card fraud detection using artificial neural network, https://doi.org/10.1016/j.gltp.2021.01.006.

[3] S. Benson Edwin Raj and A. Annie Portia, "Analysis on credit card fraud detection methods," 2011 International Conference on Computer, Communication and Electrical Technology (ICCCET), 2011, pp. 152-156, doi: 10.1109/ICCCET.2011.5762457.

[4] J. O. Awoyemi, A. O. Adetunmbi and S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis," 2017 International Conference on Computing Networking and Informatics (ICCNI), 2017, pp. 1-9, doi: 10.1109/ICCNI.2017.8123782.

[5] A. Srivastava, A. Kundu, S. Sural and A. Majumdar, "Credit Card Fraud Detection Using Hidden Markov Model," in IEEE Transactions on Dependable and Secure Computing, vol. 5, no. 1, pp. 37-48, Jan.-March 2008, doi: 10.1109/TDSC.2007.70228.

[6] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic and A. Anderla, "Credit Card Fraud Detection - Machine Learning methods," 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH), 2019, pp. 1-5, doi: 10.1109/INFOTEH.2019.8717766.

[7] A. Shen, R. Tong and Y. Deng, "Application of Classification Models on Credit Card Fraud Detection," 2007 International Conference on Service Systems and Service Management, 2007, pp. 1-4, doi: 10.1109/ICSSSM.2007.4280163.

[8] S. Makki, Z. Assaghir, Y. Taher, R. Haque, M. -S. Hacid and H. Zeineddine, "An Experimental Study With Imbalanced Classification Approaches for Credit Card Fraud Detection," in IEEE Access, vol. 7, pp. 93010-93022, 2019, doi: 10.1109/ACCESS.2019.2927266.

[9] R. Sailusha, V. Gnaneswar, R. Ramesh and G. R. Rao, "Credit Card Fraud Detection Using Machine Learning," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 2020, pp. 1264-1270, doi: 10.1109/ICICCS48265.2020.9121114.