**Fraud Detection in Banking Data by Machine Learning Technique**

**Alternative Title:**

Using Machine learning to detection in fraud in banking data with feature selection

**Aim:**

The aim is to leverage the power of machine learning to create efficient and accurate fraud detection systems that protect the interests of both financial institutions and their customers.

**Abstract:**

As technology advanced and e-commerce services expanded, credit cards became one of the most popular payment methods, resulting in an increase in the volume of banking transactions. Furthermore, the significant increase in fraud requires high banking transaction costs. As a result, detecting fraudulent activities has become a fascinating topic. In this study, we consider the use of class weight-tuning hyper- parameters to control the weight of fraudulent and legitimate transactions. We use Bayesian optimization in particular to optimize the hyper parameters while preserving practical issues such as unbalanced data. We propose weight tuning as a pre-process for unbalanced data, as well as Cat Boost and XG Boost to improve the performance of the Light GBM method by accounting for the voting mechanism. Finally, in order to improve performance even further, we use deep learning to fine-tune the hyper parameters, particularly our proposed weight-tuning one. We perform some experiments on real-world data to test the proposed methods. To better cover unbalanced datasets, we use recall-precision metrics in addition to the standard ROC-AUC. Cat Boost, Light GBM, and XG Boost are evaluated separately using a 5-fold cross-validation method. Furthermore, the majority voting ensemble learning method is used to assess the performance of the combined algorithms. Light GBM and XG Boost achieve the best level criteria of ROC-AUC = 0.95, precision 0.79, recall 0.80, F1 score 0.79, and MCC 0.79, according to the results

**Existing System:**

The proposed framework for fraud detection is presented In . As this figure shows, we first apply the desired pre-processing on the data and further divide the data into two sections: training and testing, followed by performing Bayesian optimization on the training data to find the best hyper parameters that lead to the improvement of the performance. We use the cross-validation method to obtain performance comparison in an unbalanced set and then examine the algorithms using different evaluation metrics, including accuracy, precision, recall, the Matthews correlation coefficient (MCC), the F1-score, and AUC diagrams.

**Problem Definition:**

By proposing Machine learning Algorithms, based approaches for credit card detection, such as Extreme Learning Method, Decision Tree classifier, K-Nearest neighbors, Random Forest classifier, Support Vector classifier, Logistic Regression and XG Boost .The model results leds to low accuracy.

**Proposed System:**

 Proposed approach to detecting credit card fraud using a RFE (Recursive Feature Elimination) this feature selection method used to select the best feature for a dependent variable and also we used oversampling and then using a machine learning algorithm its give a best accuracy.

**Modules:**

* Dataset Collection
* Algorithm
* Detection

**Dataset Collection:**

The credit card dataset is accessible for research purposes. The dataset holds transactions made by a cardholder over a time period, Disclosing a consumer’s transaction details is considered a problem of confidentiality, the main component analysis is applied to the majority of the dataset’s features using principal component analysis (PCA). PCA is a standard and widely used technique in the relevant literature for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss. It does so by creating new uncorrelated variables that successively maximize variance. The detail of the dataset containing time, V1, V2, V3. . . . . . V28 as PCA applied features, amount, and class labels.

**Data Preprocess:**

We propose to alter the a RFE (Recursive Feature Elimination) this feature selection method used to select the best feature for a dependent variable and also use SMOTETomke for over sampling.

**Algorithm:**

Logistic Regression, Extra Trees Classifier, Ada-Boost Classifier, Random Forest Classifier using these algorithm foe model creation we find which algorithm gives best accuracy we dump into a model file.

**Prediction:**

It's important to note that while predictions are based on available data and knowledge, they are not certainties, and unexpected events can impact their accuracy. The quality of predictions depends on the quality and relevance of the data used, the appropriateness of the models or methods employed, and the understanding of the underlying factors influencing the event being predicted.

**Hardware Requirements:**

* Hard Disk : 500GB and Above
* RAM : 4GB and Above
* Processor : I3 and Above

**Software Requirements:**

* Operating System : Windows 10 (64 bit)
* Software : Python-3.6.3
* Tools : Anaconda

**CONCLUSION AND FUTURE WORK**

In this paper, we studied the credit card fraud detection problem in real unbalanced datasets. We proposed a machine- learning approach to improve the performance of fraud detection. We used a publicly available ‘‘credit card’’ dataset with 28 features and 0.17 percent of the fraud data. We proposed two methods. In the proposed Light GBM, we used class weight tuning to choose the proper hyper parameters. We used the common evaluation metrics, including accuracy, precision, recall, F1-score, and AUC. Our experimental results showed that the proposed Light GBM method improved the fraud detection cases by 50% and the F1-score

by 20% compared with the recently presented method in .We improve the performance of the algorithm with the help of the majority voting algorithm. We also improved the criteria by using the deep learning method. The assurance of the results of MCC for unbalanced data proved that, compared to other criteria of evaluation, it’s stronger. In this paper, by combining the Light GBM and XG Boost methods, we obtained 0.79 and 0.81 for the deep learning method. Using hyper parameters to address data unbalance compared to sampling methods, in addition to reducing memory and time needed to evaluate algorithms, also has better results. For future studies and work, we propose using other hybrid models as well as working specifically in the field Cat Boost by changing more hyper parameters, especially the hyper parameter number of trees. Also, due to hardware limitations in this study, the use of stronger and better hardware may bring better results that can ultimately be compared with the results of this study

**Archiecture diagram:**

Pre-processed Dataset

Feature Extraction

Machine learning alorithm

Real time

User input

User interface

Prediction

Model compile

 Raw dataset

Trained Model