Detection Of Illegitimate Credit Card Transactions Using Machine Learning

Dhananjeyan M S
Department of Computer Science and Engineering
Meenakshi Sundararajan Engineering College
Chennai,India
msdhanan99@gmail.com

Raguram K B
Department of Computer Science and Engineering
Meenakshi Sundararajan Engineering College
Chennai,India
raguram2684@gmail.com

Gokul R
Department of Computer Science and Engineering
Meenakshi Sundararajan Engineering College
Chennai,India
gokul.agnschool@gmail.com

Dr. M K Sandhya
Professor
Department of Computer Science and Engineering
Meenakshi Sundararajan Engineering College
Chennai,India
mksans@gmail.com

Abstract: **In the present time, online installment framework and net banking are broadly acknowledged everywhere throughout the world as they make the exchanges a lot simpler and quicker. Visa exchanges are quickly expanding throughout the years that lead to a higher pace of online fraud and resulting misfortunes by banks just as purchasers. Lawbreakers can utilize a few innovations, for example, Trojan or Phishing to take the data of others' credit cards. Existing systems recognize even the legitimate transactions as fake ones. Hence a successful strategy to recognize fake exchanges is significant. Right now we give the answer for this issue by Machine Learning and with the assistance of calculations, for example, SMOTE and Random Forest. To test this model a few assessment parameters, for example, AUPRC, AUROC, precision score, recall score and F1 score are determined. The accuracy score of this model is 99%.**

I. INTRODUCTION

In the present time, online installment framework and net banking are broadly acknowledged everywhere throughout the world as they make the exchanges a lot simpler and quicker. Credit card exchanges are quickly expanding throughout the years that lead to a higher pace of online data fraud and ensuing misfortunes by banks just as shoppers. Hoodlums can utilize a few innovations, for example, Trojan or Phishing to take the data of others' charge cards. Run of the mill association loses about 5% of its income because of misrepresentation exchanges. There are different models which are utilized for distinguishing the misrepresentation exchanges dependent on the conduct of the exchanges and these strategies can be delegated two general classes, for example, managed learning and unaided learning calculation. The current framework utilizes neighborhood anomaly factor and disengagement timberland strategies to distinguish the fake exchanges however this delivers an exactness up to 97% as it were. The point of this paper is to improve the precision of finding the fake exchanges.

II. EXISTING SYSTEM

In the current framework [6], a relative investigation of local outlier factor and isolation forest techniques are given to recognize the false exchanges. This model attempts to get high extortion exchange inclusion at exceptionally low false alert rate and taking care of huge volumes of exchanges, henceforth giving a technique to identify credit card cheats and giving outcomes in less time. Utilizing this model clients exchange design is broke down and any deviation from ordinary example is considered as fake exchange. It makes detection handling very easy and tries to eliminate the complexity. The data set for this paper depends on genuine value-based information by a huge European organization and individual subtleties in information are kept classified. Exactness of this model utilizing nearby anomaly factor is 97% and utilizing isolation forest is 76%. The accuracy of the outcomes got from these strategies are less when compared with the proposed framework.

III. LITERATURE SURVEY

The paper [1] utilizes a developmental Simulated Annealing calculation to prepare the Neural Networks for Credit Card extortion identification in real-time scenario. Artificial Neural system is utilized in this paper to distinguish the sort of exchanges and the upsides of doing so are likely gainful for the associations and for singular clients regarding cost and time productivity. Hyperbolic tangent activation curve is utilized to discover the similarities. The drawback of this is a few lawful clients are misclassified as fake clients.

On other turn in paper [2] KNN calculation and outlier discovery techniques are executed to upgrade the best answer for the misrepresentation location issue. These methodologies are refuted to limit the caution rates and increment the extortion discovery rate. The preferred position is that there is no necessity of predictive model before arrangement. Just depends on neighbors for the order. In any case, the accuracy of the framework relies upon the separation of the neighbors.

The paper [3] examines the exhibition of naive bayes, k-nearest neighbor and logistic regression on exceptionally skewed credit card misrepresentation information. A cross breed system of under-sampling and oversampling is completed on the skewed information. The favorable position is that the presentation of the methods is assessed dependent on accuracy, sensitivity, specificity, precision, Matthews correlation coefficient and balanced classification rate. Complexity of examination has expanded. In any case, execution of classifiers fluctuates across various assessment measurements.

In paper [4] Support Vector Machines are utilized to distinguish the extortion discoveries and decrease the misrepresentation cautions. This technique is fit for distinguishing the fake exchange on schedule. Be that as it may, not Dependable and can't recognize misrepresentation exchanges now and then.

IV. PROPOSED SYSTEM

The dataset utilized in this paper is from kaggle and this dataset contains **284,807** exchanges that occurred during September 2013 at Europe. Since it comprises of exchange subtleties of the clients out of the 30 characteristics in the dataset just 2 are noticeable parameters specifically Amount and Time. Remaining 28 characteristics are in PCA format with the goal that estimations of it are obscure and kept secret. In this dataset there is an extra factor called class variable which has a value either 1 or 0. In the event that the value is 0, at that point it demonstrates it is a legitimate exchange else it is a false exchange.

We standardize the dataset by reshaping the values in the Amount section for calculation purposes. We also come to a conclusion that time of the transaction cannot help to distinguish the type of the transaction and hence two columns Time and Amount are dropped from the dataset and a new column is added where the values of normalized amount are alone present.

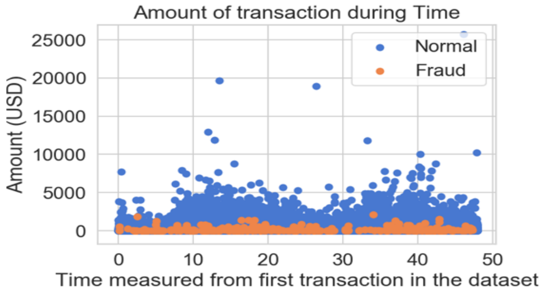


Fig 1:Scattered graph(time versus amount)

We arrange the dataset into three to be specific train, validation and test. The train dataset will be **70**% and test will be **30**% of the dataset. The validation dataset is a subset of train dataset and it is **10**% of the train dataset. For arrangement of the dataset we utilize a product called "Scikit-learn". Scikit-learn is a free programming for machine learning library in python where it contains highlights like Classification, Regression, Clustering calculations and furthermore different calculations to interoperate with Python.

A) SMOTE (Synthetic Minority Over Sampling Technique)

In this dataset the ratio of typical exchanges to that of fake exchanges is little. The count of false exchanges present in the whole dataset is 492 and the remaining are lawful exchanges. Therefore, a class imbalance issue happens. If training takes place without addressing this problem then the outcomes of this model would not be effective. So as to defeat this issue we use SMOTE. This is an oversampling method which assists with bringing the proportion of legitimate exchanges and false exchanges in the range 1:1. In the wake of applying SMOTE the level of fake and lawful exchanges are half separated and before applying SMOTE the level of the false exchanges and legitimate exchanges were 0.17% and 99.73% individually. With the usage of Smote the size of the train dataset doubles in size and now the train dataset is balanced.

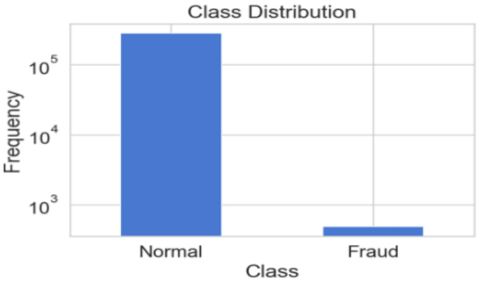


Fig 2: Before SMOTE analysis

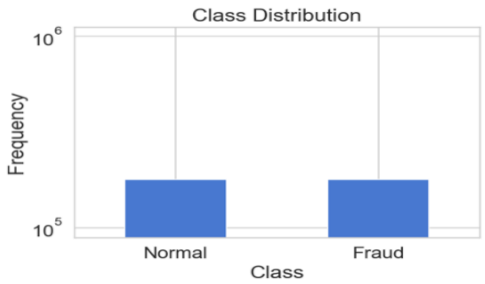


Fig 3: After SMOTE analysis

B) Random Forest Algorithm

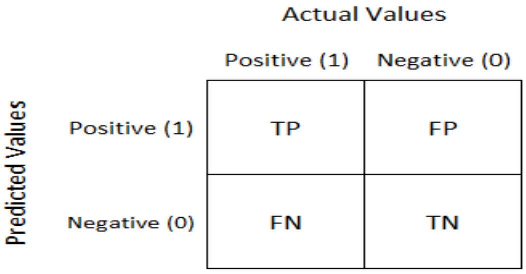
When the train dataset is adjusted subsequent to applying Smote the following stage is preparing the dataset. This calculation develops decision trees to recognize the sort of exchange and a confusion matrix is created with the assistance of this calculation. Each decision tree delivers up to some condition. Finally voting process takes place to find out the common type of classification [5]. The upside of utilizing this calculation is, this can be utilized for characterization just as for regression. The base accuracy ensured with this calculation is 90%. The confusion matrix returns four qualities, for example, true positive, true negative, false positive, false negative. In view of these qualities' assessment parameters are determined and proficiency of the model is tested.

V. EVALUATION RESULTS

Precision of a calculation can't the only one be a central factor for making a decision about the technique. We additionally compute the accompanying measures, for example,

- Precision
- Recall
- F1 Score
- Support
- AUPRC
- AUROC

These all standard of correctness are depend upon the Actual and Predict class, so we draw a 2×2 confusion matrix



Predicted Values \ Actual Values	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

Fig 4:Format of confusion matrix

CONFUSION MATRIX
[85284 23]
[23 113]

- *Evaluated Confusion Matrix*

i. *True Positive (TP):* These values arecorrectively predicted positive that means value of both actual class and predicted class are YES.

ii. *True Negative (TN):* These values are correctively predicted Negative that means value of both actual class and predicted class are NO.

iii. *False Positive (FP):* when value of actual class is NO and value f predicted class is YES.

iv. *False Negative (FN):* when value of actual class is YES and value f predicted class is NO.

False Positive and False Negative classes occur when actual class contradicts with predicted class.

v. *Precision*: It is Ratio of correctly predicted Positive observations to the Predicted positive observations.

Equation 1:

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

vi. *Recall*: It is ratio of correctly predicted positive observations to the all observations in actual class YES.

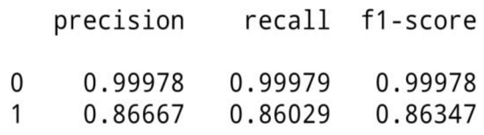
Equation 2:

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

vii. *F1 Score*: It is the weighted average of Precision and Recall. Therefore this score takes both false negatives and false positives into account.

Equation 3:

$$\text{F1} = 2 \times \frac{Precision * Recall}{Precision + Recall}$$



```
       precision    recall  f1-score

0        0.99978   0.99979   0.99978
1        0.86667   0.86029   0.86347
```

Fig 5: Evaluated parameters for class 0 and 1

AUPRC:

The Area Under the Precision-Recall Curve (AUPRC) is another exhibition metric that you can use to assess a classification model. If your model achieves a perfect AUPRC, it means your model can find all of the positive samples (perfect recall) without accidentally marking any negative samples as positive (perfect precision.).

The Proposed system Scores **83.6**%.

AUROC:

AUC - ROC a performance measurement for classification problem at different thresholds settings. ROC is a probability curve and AUC speaks to degree or proportion of detachability. It tells how much model is equipped for recognizing classes. Higher the AUC, better the model is at foreseeing 0s as 0s and 1s as 1s

- 90-1 = excellent (A)
- 80-90 = good (B)
- 70-80 = fair (C)
- 60-70 = poor (D)
- 50-60 = fail (F)

The proposed system scores **95** in performance.

VI. CONCLUSION

Credit card swindles keeps growing at an upsetting rate. In spite of the fact that numerous investigations and correlation with accessible strategies, our proposed framework shows broadened execution including both processing quality and time utilization with an exactness pace of **99.97**%.

REFERENCES

[1]. AzeemUsh Shan Khan, Nadeem Akhtar and Mohammad Naved Qureshi-" Real-Time Credit-Card Fraud Detection using Artificial Neural Network Tuned by Simulated Annealing Algorithm"ACEEE Proc. of Int. Conf. on Recent Trends in Information, Telecommunication and Computing, ITC 2014.

[2]. N.MaliniM.Phil , Dr.M.Pushpa Assistant Professor, "Analysis on Credit Card Fraud Identification Techniques based on KNN and Outlier Detection" 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)

[3]. John O. Awoyemi, Adebayo O. Adetunmbi, Samuel A. Oluwadare - "Credit card fraud detection using Machine Learning Techniques: A Comparative Analysis", 2017 International Conference on Computing Networking and Informatics (ICCNI)

[4]. Nancy Demla ,Alankrita Aggarwal, "Credit Card Fraud Detection using SVM and Reduction of False Alarms" , 2016 International Journal of Innovations in Engineering and Technology (IJIET).

[5]. M.Suresh Kumar, V.Soundarya , E.Aswini , E.S.Keerthika-"Credit Card Fraud Deteection Using Random Forest Algorithm", 3rd Int. Conf. on Computing and Communication Tecnologies ICCCT 2019

[6]. Hyder John, Sameena Naaz- "Credit Card Fraud Detection Using Local Outlier Factor and Isolation Forest", Intl. Journal of Computer Sciences And Engineering 2019