The Role of Logistic Regression and Local Outlier Factor for Credit Card Fraud Identification

[1]Kanasani Nagamani, [2]Sateesh Enapakurthi

[1] Student, [2] Assistant Professor
Amrita Sai Institute of Science and Technology
Autonomous NAAC with A Grade, Andhra Pradesh, India

***Abstract*-Credit cards are a crucial financial tool that enables its users to make purchases and pay at a later date. Issued by financial customs, credit cards give users a pre-agreed credit limit that they can use for their purchases. MasterCard extortion is a type of datafraud where crooks make buys utilising a Visa account which doesn't have a place with them. The two primary tactics for reducing frauds and losses caused by fraudulent conduct are fraud detection systems and fraud prevention. Frauddetection is tracking the behaviours of large groups of people in order to estimate, perceive, or identify obnoxious activity, such as fraud, intrusion, or defaulting. The Local Outlier Factor is a technique for detecting aberrant datapoints by comparing a data point's local variability to that of its neighbours. Under the Supervised Learningapproach, one of the most prominent Machine Learning algorithms is logistic regression.**

Keywords: Local Outlier Factor, Logistic Regression, Fraud detection

I. INTRODUCTION

Charge cards have been utilized in individuals' daily existence for purchasingitems and administrations where buying can be disconnected as well as on the web. Advancement in data innovation and upgrades in correspondence channels, extortion is spreading everywhere, bringing about gigantic monetary misfortunes. Misrepresentation can be characterized as a situation where an individual purposes another person's Visa for ill-conceived exercises while the proprietor and the card- giving specialists know nothing about theway that the card is being utilized. The two primary components to keep away from fakes and misfortunes because of false exercises are extortion recognizable proof frameworks and misrepresentation anticipation. Misrepresentation ID includes observing the exercises of populaces of clients to appraise, see or distinguish frightful way of behaving, which comprises of extortion, interruption, and defaulting. The calculations we utilized in this recognizable proof are Logistic Regression and Local Outlier Factor. Nearby Outlier Factor is a calculation utilized for observing strange pieces of information by assessing the nearby changeability of a given data of interest in contrast with its neighbors.
Calculated relapse is one of the most famousMachine Learning calculations, which goes under the Supervised Learning procedure. It is likewise utilized for taking care of grouping issues.

II. MOTIVATION

A. Literature Survey

Anjali Singh Rathore; Ankit Kumar; Depanshi Tomar; Vasudha Goyal; Kaamya Sarda -Credit Card Fraud Detection usingMachine Learning(IEEE-2021).On severely unbalanced data, this research evaluated the

performance of Decision Tree, RandomForest, K-nearest neighbours, and Logistic regression.[1]

Trivedi and M. Mridushi—Credit Card Fraud Identification, Ijarccen. In this Thepaper employs heredity computations thatinclude algorithms for predicting the optimalsolution to a problem and deleting unspoken discoveries from phoney interactions. Extortion differentiating evidence is a primary goal of hereditary models.[2]

R. Banerjee, G. Bourla, S. Chen, S. Purohit, and J. Battaglia—Credit Card Fraud Identification as a Comparative Analysis of Machine Learning Algorithms The authorsof this research described the best computational technique and the best-performing combination of criteria for detecting Visa misrepresentation.[3]

D. Tripathi, T. Solitary, Y. Sharma, and S. Dwivedi, —Detection of credit card fraudusing the Local Outlier Factor. The idea of leveraging the Local Outlier Factor to combat faking recognition both fordisconnected and online purchases using MATLAB and the instalment number as the misrepresentation test is presented in this paper. [4]

C. P. Lim, M. Seera, A. K. Nandi, K. Randhawa, and C. K. Loo—Credit Card Fraud Detection Using AdaBoost and Majority Voting‖IEEE Access. Defaultmodels such as NB, SVM, and DL, as well as cutting-edge AI models such as Ada Boost, were used in this paper to determine how to deal with charge card deception.[5]

III. METHODOLOGY

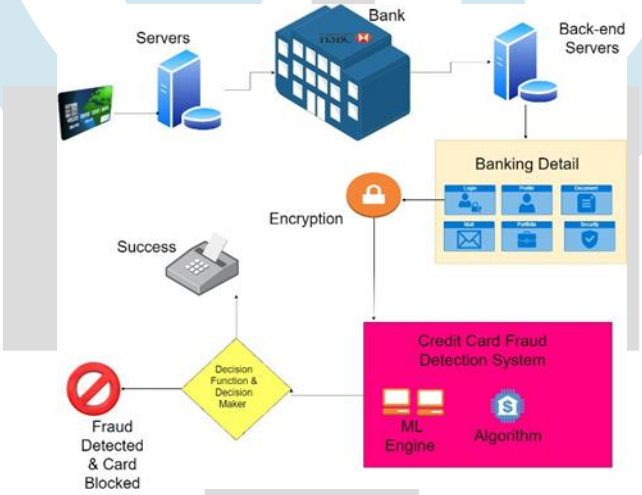

Fig .1. System Architecture

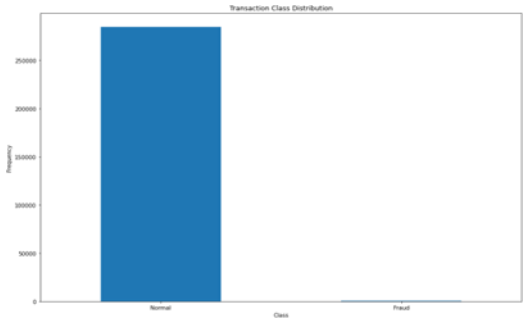

Fig .2. Transaction Class Distribution

The above graph demonstrations that the number of deception transactions is extremely low when equaled to legal transactions. Hence the fact that the dataset ishighly unbalanced is inferred.

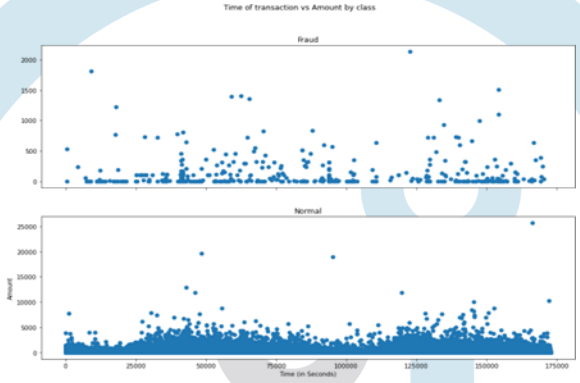

Fig .3. Scatter plot of Transaction Time vs TransactionAmount

The above scatter plot shows the amount of transactions with respect to time. The first plot corresponds to Fraud transactions while the second corresponds to Legaltransactions.

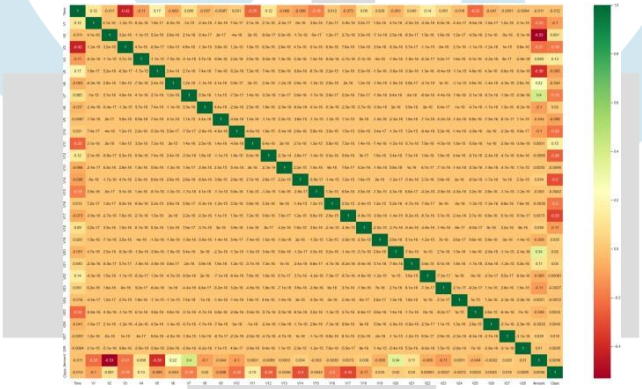

Fig4: Correlation among the features

The above figure is the hotness map, which depicts the relationship among the gauging factors and class variable. We can construe that include V7 with sum and V20 with sum showed high connection between's them.

Logistic Regression:

It is an Supervised learning grouping calculation.

In 1972, Nelder and Wedderburn proposed this model with a work to give a method for utilizing straight relapse to the issues which were not straightforwardly appropriate forapplication.

In Linear relapse the ideal output,which is the reliant variable, is dichotomous, and thatimplies it can have just two potential values.It is signified by Y.

The confounders are known as theautonomous factors indicated by X.Thereare two problems with supervised learning 1.Unbalanced class sizes

In credit card transactions, the amount offraud businesses is way too a smalleramount than the amount of

legaltransactions.

2. Mislabeled data
Some of the fraud transactions may go undetected; leading mislabeled the transactions thus miscalculating the output

Local Outlier Factor (LOF):
(LOF) is an unsupervised anomaly detectionalgorithm proposed by Markus M. Breunig, Hans Peter Kreigel and Jorg Sander. This method focuses on computing local density deviation of a data point to its neighbours. On comparing the data point by means of local densities of its neighbours, regions with similar density and regions with lower densities can be identified. The data point is considered as an outlier if and only if its density is comparatively low to its neighbours.
The dataset used in this solution has fraudulent transactions of a very smallsize(492 transactions) when compared to genuine transactions(2 lakh transactions).This factor supports the Local outlier Factor algorithm to be a suitable method in detecting fraudulent transactions.

Data sets:
The dataset is obtained from kaggle which isin csv format (Comma Separated Values).
This dataset contains the credit card transactions that took place in the timelineof 48 hours in September 2013, Europe. In total it has 31 features out of which 28 features remain anonymous and named asv1-v28. These 28 features are madeanonymous by transforming the values usingPCA(Principal Component analysis) due to confidentiality. The remaining three featurescorrespond to Time, Amount and Class. Time defines the amount of time between two consecutive transactions. Amount defines the amount of money handled while Class defines the class of a transaction. The class feature holds two labels, namely, 0 and
1. 0 indicates that the transaction is genuine,whereas, 1 indicates that the transaction is fraudulent.
This dataset is highly unbalanced; it has only 492 fraud transactions out of 2, 84,807 transactions.

Tools:

The lists of tools used in this analysis are as follows:

- This anticipated model is employed in Python.
- For simpler tasks such as data storage and Transformation, NumPy and Pandas are used.
- Intended for data analysis and visualization, Matplotlib is used.
- Seaborn is used for statistical datavisualization and for algorithms we used Sklearn.

IV. EXPERIMENTAL RESULTS

The consequences of the proposed CreditCard Fraud Identification utilizing Machine Learning are portrayed here.

A. Results

Though the algorithms work fine, the collection of data is a huge task. Collecting data requires cooperation with banks and itis quite difficult as the data to be shared is highly sensitive. The data list that has been shared even after undergoing many transformations to increase confidentiality has to call off their cards to circumvent any supplementary risk.

```
LogisticRegression(max_iter=1000): 10
Accuracy Score :
0.949238578680203
Classification Report :
              precision    recall  f1-score   support

           0       0.93      0.97      0.95        99
           1       0.97      0.93      0.95        98

    accuracy                           0.95       197
   macro avg       0.95      0.95      0.95       197
weighted avg       0.95      0.95      0.95       197
```

```
LocalOutlierFactor(contamination=0.0017234102419808666): 97
Accuracy Score :
0.9965942207085425
Classification Report :
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     28432
           1       0.02      0.02      0.02        49

    accuracy                           1.00     28481
   macro avg       0.51      0.51      0.51     28481
weighted avg       1.00      1.00      1.00     28481
```

B. Interpretation:

- Logistic regression is found to have an accuracy of 94.92% while its errors are found to be 10. This is due to the unbalanced data set which is balanced by taking data members from the dataset such that the fraudulent and legal transactions are of the same size.
- The accuracy of Local outlier factor is found to be 99.65% whereas the number of errors are 97.
- From this analysis we can infer that the performance of the Local outlier factor model is comparatively high with logistic regression.

C. Comparative Analysis:

The performance of the proposed models areanalyzed and compared with respect to their accuracy.

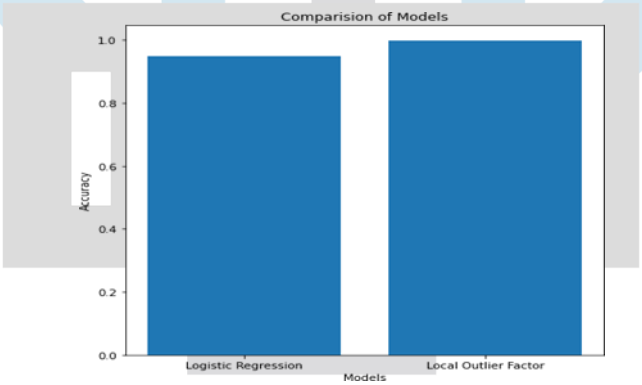

Fig 5 Comparison of Logistic Regression and Local Outlier Factor

CONCLUSION AND FUTURE SCOPE

It is fundamental for Mastercardorganizations to have the option to perceive deceitful Visa exchanges. With the developing utilization of Visas for buys, the dangers of Mastercard cheats develop rising fundamentally. In this paper an examination of Mastercard misrepresentationdistinguishing proof was portrayed on a freely accessible dataset using MachineLearning procedures, for example, Local anomaly factor and Logistic Regression. While we were unable to connect objective of 100 percent exactnessin extortion ID, we wound up making a framework that can, with sufficient opportunity and information, get exceptionally near that goal.The very nature of this undertaking takes into account numerous calculations to be coordinated together as modules and their outcomes can be consolidated to build the precision of the end-product.

REFERENCES

[1] R. Banerjee, G. Bourla, S. Chen, S. Purohit, and J. Battipaglia, "Comparative Analysis of Machine Learning Algorithms through Credit Card Fraud Detection," pp. 1–10, 2018.

[2] T. Patel and M. O. Kale, "A Secured Approach to Credit Card Fraud Detection Using Hidden Markov Model," vol. 3, no. 5, pp. 1576–1583, 2014.

[3] Hobson, A.The Oxford Dictionary of Difficult Words. The Oxford University Press, New York (2004)

[4] Bolton, R. J., Hand, D. J.: Statistical fraud detection: A review. Statistical Science, Vol. 17, No. 3 (Aug., 2002), pp. 235-249 (15 pages)