Credit Card Fraud Detection Using CNN

1Mrs.M.Madhavi, 2K.R.Venkat Reddy, 3Burra Swetha,4R.Badri Kumar

1Assistant Professor, Department of Artificial Intelligence, Anurag Group Of Institutions,
234Student,Department of Artificial Intelligence, Anurag Group Of Institutions, Hyderabad, India,

Abstract—This paper talks about a strategy within the fraud detection interface region. The approach proposed is to utilize imbalanced profoundly skewed value-based information and a convolutional organize for the discovery of fakes. The dataset utilized here is the machine learning kaggle dataset for credit card extortion discovery that contains profoundly skewed information. The assessed highlights are 1 for fraud and for non-fraud class. The examination of extortion discovery was an vital device in keeping money divisions. These days, the counterfeit neural organize has ended up the slightest effective strategy for credit card extortion discovery. The framework right now utilized to distinguish extortion is tormented by misclassifications and exceedingly wrong positives. In such situations here this term paper employs the in participation of convolutional neural arrange layers in an endeavor to construct a show for identifying credit card extortion that gives us a tallevel of exactness.

Keywords: Convolutional Neural Network Layers, Imbalanced dataset, SMOTE, Dropout, Credit Card

I. INTRODUCTION
Credit card alludes to a card that's more often than not issued to a client (cardholder) and ordinarily permits a client to buy products or administrations or pull back cash in development inside their credit restrain. Credit cards give cardholders with a period of time, i.e., Clients can move on to the following charging cycle and pay their cash afterward in a endorsed time. Credit card extortion is simple prey. Risk-free costs can be pulled back in a brief sum of time without the owner's information. In today's world, we are on the express prepare to a cashless society. In 2016 add up to of non-cash exchanges expanded by 10.1% from 2015 for a addup to of 482.6 billion exchanges. In 2017, there were 1.579 information breaches and almost 179 million records, of which 133,015 were detailed on credit card extortion. On the flip-side false exchanges are on the rise as well. This must be halted as soon as conceivable. Because it advances, banks turn to EMV cards. It's a savvy card that stores information in a compact circuit instead ofa attractive strip, making certain card installments more secure, but still taking off card-not-present fakes at higher rates.Criminalshave moved their center on exercises related to CNP exchanges as the security of chip cards have increased.

The notoriety of on-line shopping has developed day by day in today's world. Concurring to a 2005 ACNielsen study, 1 in 10 of the world's populace buys online. These days, credit card is the foremost prevalent strategy of installment. As the number of creditcard clients increments around the world, character robbery is on the rise and extortion is on the rise. Credit card-based buys can be partitioned into two sorts:
1) physical card buys and 2) visual card buys. At the buy of a physical card, the cardholder personally presents the card to create a payment. Whereas buying a physical card, the assailant has to take the credit card and construct a signature to purchase it. For visual card buys, as it were card subtle elements are required such as card number, termination date, secure code, etc. Such buys are as a rule made online or over the phone. To commit extortion in these sorts of buys, a individual essentially should know the points of interest of his or her card. The online shopping mode is made for the most part by credit card. Credit card extortion has been expanding day by day. The sum of budgetary misfortunes due to credit card extortion increments as the utilize of credit card gets to be more common. Security implies utilizing your credit card securely and maintaining a strategic distance from false appearances. The reason of security is to anticipate the utilize of false credit cards. In cases of extortion there are issues such as misplaced cards, misplaced cards, stolen cards, application extortion, false mail, postal extortion and unpaid extortion (NRI). To decrease this chance, credit card security is required.

II. LITERATURE SURVEY

Fraud in any way may be a criminal action and is an offense, credit card extortion is taking cash. There are numerous considers in which they attempted to discover whether a exchange is fraud or not. Still having numerous challenges and tries to overcome those issues Firstly, numerous utilized Information Mining Strategies to discover false exchanges by utilizing a few Conventional approach, which isn't customary and these days fraudsters are so shrewd that they can do fraud without damaging rules so, Utilizing Machine learning is required.

Machine learning moreover comes with challenges. So, here a intensely unbalanced data (Data set from Kaggle by mlg-ubj) set is considered, so that it'll give us the leading calculation to utilize in conjunction with its challenges. Because it is overwhelming imbalanced, even if the proposed calculation is sweet or not, it gives us an exactness of approximately 99.9%. So, here the beneath examining is considered to supply us with great comes about as in Exception location and evacuation calculations are utilized to precisely foresee false exchanges of a credit card exchange dataset.

© 2023 IJRTI | Volume 8, Issue 4 | ISSN: 2456-3315
Modern strategies for credit card extortion location with a part of investigate strategies and a few extortion discovery methods with a extraordinary intrigued within the neural systems, information mining, and dispersed information mining. Numerous other strategies are utilized to identify such credit card extortion. When done the writing study on different strategies of credit card extortion discovery, ready to conclude that to identify credit card extortion there are numerous other approaches in Machine Learning itself. The inquire about on credit card extortion location employments both Machine Learning and Profound Learning calculations. In this area, we upgrade the work wiped out two different points:

(i) the strategies that are promptly accessible for extortion discovery, and

(ii) The procedures that are accessible to handle the imbalanced information. To handle the imbalanced information A few ofthe procedures are accessible. They are (a)classification strategies (b) examining strategies (c) taking after methods. Here are a few of the Machine Learning calculations that are utilized for credit extortion discovery are back vector machine(SVM), choice trees, calculated relapse, slope boosting, K-nearest neighbor, etc;

In 2019, Yashvi Jain, Namrata Tiwari, Shripriya Dubey, Sarika jain have inquired about different strategies for credit cards extortion location such as bolster vector machines(SVM), Artificial neural networks(ANN), Bayesian Systems, Hidden Markov Show, K-Nearest Neighbors (KNN), Fluffy Rationale framework and Choice Trees. In their paper, they have watched that the calculations k-nearest neighbor, decision trees, and the SVM grant a medium level exactness. The Fluffy Rationale and Calculated Relapse grant the least precision among all the other calculations. Neural Systems, gullible bays, Fluffy frameworks, and KNN offer a high detection rate. The Logistic Regression, SVM, choice trees offer a tall discovery rate at the medium level. There are two calculations to be specific ANN and the Naïve Bayesian Systems which perform superior at all parameters. These are exceptionally much costly to train. There is a major downside in all the calculations. The downside is that these calculations do not allow the same result in all sorts of situations. They grant superior comes about with one sort of datasets and destitute comes about with another sort of dataset. Calculations like KNN and SVM grant great comes about with small datasets and calculations like calculated relapse and Fluffy rationale frameworks donate great precision with crude and unsampled information.

In 2019, Heta Naik, Prashasti Kanikar, has done their investigate on different calculations [4] like Naïve Bayes, Calculated Relapse, J48, and Adaboost. Naïve Bayes on among the classification calculation. This calculation depends upon Bayes hypothesis. Bayes's hypothesis finds the likelihood of an occasion that's happening is given. The Calculated relapse calculation is comparative to the direct relapse calculation. The direct relapse is utilized for the forecast or estimating the values. The calculated relapse is for the most part utilized for the classification errand. The J48 calculation is utilized to create a choice tree and is utilized for the classification issue. The J48 is the expansion of the ID3 (Iterative Dichotomies). J48 is one of the foremost broadly utilized and broadly analyzed ranges in Machine Learning. This calculation basically works on steady and categorical factors. Adaboost is one of the foremost broadly utilized machine learning calculations and is primarily created for binary classification. The calculation is mainly utilized to boost the execution of the decision tree. Typically moreover primarily utilized for the classification of the relapse. The Adaboost calculation is extortion cases to classify the exchanges which are extortion and non-fraud. From their work they have concluded that the most noteworthy exactness is gotten for both the Adaboost and Logistic Regression. As they have the same precision the time calculate is considered to select the leading calculation. By considering the time calculate they concluded that the Adaboost calculation works well to distinguish credit card fraud.

In 2019 Sahayasakila V, D.Kavya Monisha, Aishwarya, Sikkakoli Venkatavisalakshiswshai Yasaswi have clarified the Two vital algorithmic procedures which are the Whale Optimization Procedures (WOA) and Destroyed (Manufactured Minority Oversampling Methods). They basically pointed to progress the joining speed and to fathom the information lopsidedness issue. The course lopsidedness issue is overcome utilizing the Destroyed strategy and the WOA strategy. The Destroyed procedure segregates all the exchanges which are synthesized are once more re-sampled to check the information exactness and are optimized utilizing the WOA procedure. The calculation moreover moves forward the joining speed, reliability, and productivity of the framework.

In 2018 Navanushu Khare and Saad Yunus Sait have clarified their work on choice trees, arbitrary timberland, SVM, and calculated relapse. They have taken the profoundly skewed dataset and worked on such sort of dataset. The execution assessment is based on precision, affectability, specificity, and accuracy. The comes about demonstrate that the precision for the Calculated Relapse is 97.7%, for Choice Trees is 95.5%, for Irregular Timberland is 98.6%, for SVM classifier is 97.5%. They have concluded that the Arbitrary Timberland calculation has the most elevated exactness among the other calculations and is considered as the finest calculation to identify the extortion. They also concluded that the SVM calculation includes a information lopsidedness issue and does not grant superior comes about to distinguish credit card extortion.

III. PROBLEM STATEMENT
The main problem is online payment doesn’t even require the presence of real card. Any person with the card details can makesuch crime (fraudulent transactions). Card holder comes to know about such fraud act, only when the fraud transaction has been occurred. The sole aim of this project is to develop Credit Card Fraud Detection System using Machine Learning.
3.1 Existing Systems
Credit card fraud detection requires information examination and behaviors observing. In case of the existing framework, the fraud is identified once the extortion is done that’s, the fraud is recognized after the complaint of the card holder. And in this way, the card holder faces a part of inconvenience some time recently the investigation wrap up. Additionally, as all the exchange is kept up in a log, so it requires the support of gigantic information. These days part of online buy are made so it doesn't know the individual who is utilizing the card online for exchange and fair capture the IP address for confirmation reason. In this manner, there needs a offer assistance from the cyber wrong doing to investigate the fraud. To dodge the complete over cons the framework is made to distinguish the fraud in a best and simple way

As efficient as the credit card fraud detection is, it fails when inclusivity at somepoint. We recognize the limitations faced by the existing framework: Computational performance: As a high-cost problem it is desirable for financial fraud to be detected immediately to avoid advance delay. Exceptionally small research has been conducted on the computational performance of fraud detection methods for utilize in real-time situations. Evolving problem: Fraudsters are continually modifying their techniques to remain undetected. As such detection methods are required to be able to constantly adapt to unused fraud techniques to help people.Disproportionate misclassification costs: Fraud detection is primarily a classification problem with a vast difference in misclassification costs. Research on the performance of detection methods with respect to this factor is an area which needs encourage attention. To overcome all the above points, we proposed a modern framework. This will reduce the illegal transactions being done and also guarantee safety to the people related to his financial matters

3.2 Proposed System
The main Aim of the project is to detect the whether the given transaction is a genuine transaction or a fraud transaction by using the CNN. The Main Advantage of using CNN is because of the convolution layer which Extracts all important features. Our approach for Credit Card Fraud Detection contains of Machine Learning and Deep learning techniques. They contain different algorithms to train a model and to test that model

Step 1 – Collection of the data from different sites like kaggle. Step 2 – Applying the Data Pre-processing steps. Step 3 – Applying the SMOTE Oversampling technique. Step 4 – Applying the CNN Algorithm. Step 5 – Calculate score to get a good model.

IV. DATASET
The dataset contains cardholder’s activities for two days (Two days within the month of September). There are 284,807 exchanges, in them there are 492 false exchanges. The dataset contains a add up to of 31 traits i.e 31 columns from them one of the property Time taken to prepare the total exchange in seconds following one is the sum which is utilized in that exchange, the course which determines whether the exchange may be a veritable exchange or an extortion exchange and the remaining properties are V1, V2, V3, ..., V28 these properties are covered up to secure the security of the clients. They more often than not contain Title of Card Holder, put of exchange, stage utilized for making exchange etc. The database is profoundly lopsided (unbalanced or Overfitted)

V. PROPOSED METHODOLOGY
Step 1 – Collection of the data from different sites like kaggle. Step 2 – Applying the Data Pre-processing steps. Step 3 – Applying the SMOTE Oversampling technique. Step 4 – Applying the CNN Algorithm. Step 5 – Calculate score to get a good model.
5.1. Data Collection

Data collection or gathering is the process of gathering and measuring information on targeted variables in an established system, which then enables one to answer relevant questions and evaluate outcomes. Data collection is a research component in all study fields, including physical and social sciences, humanities, and business. While methods vary by discipline, the emphasis on ensuring accurate and honest collection remains the same. The goal for all data collection is to capture quality evidence that allows analysis to lead to the formulation of convincing and credible answers to the questions that have been posed. Data collection and validation consists of four steps when it involves taking a census and seven steps when it involves sampling. The data is taken from the Kaggle it contains 284,807 transactions and 31 attributes from them some of them are amount, class, time taken for transaction and the remaining 28 attributes are the hidden attributes which are hidden from people to maintain the privacy of the users.

5.2. Data Preprocessing and cleaning

Information preprocessing can allude to control or dropping of information some time recently it is utilized in arrange to guarantee or upgrade performance and is an imperative step within the information mining prepare. The express "waste in, waste out" is especially appropriate to information mining and machine learning ventures. Data-gathering strategies are frequently freely controlled, coming about in out-of-range values (e.g., Pay=−100), incomprehensible information combinations (e.g., Sex: Male, Pregnant :Yes), and lost values, etc. The collected information may contain a few the invalid esteem. When the dataset was being arranged some of the values within the transactions may be misplaced so when we are attempting to fit our show on the dataset which have the invalid values at that point our demonstrate will toss an mistake saying that the a few of the values which are show in our collected dataset are null(empty) so we must clear those specific exchange which means we must drop those specific exchanges in arrange to dodge the mistakes.

5.3. Applying Oversampling Techniques (SMOTE)

In the dataset which we collected there are only 492 fraudulent transactions among the 284,807 when we use this dataset then our model will not be able to predict all the different cases because of the over-fitting problem. The genuine or the true transactions are more in number so there will not be many cases for the model to predict the fraudulent transactions and so when we use the SMOTE algorithm then the number of the Genuine and the Fraudulent transitions will be made equal. What the SMOTE Algorithm does is it tries to replicate the type of the class which is having less number of values and tries to remove the problem of the oversampling and when the sample is picked then the number probability for the 2 classes to be picked will be the same.
Information is at the heart of each ML issue. Without appropriate information, ML models are similar to bodies without soul. But in today's world of 'big data' collecting information isn't a major issue any longer. We are intentionally or unknowingly creating gigantic datasets each day. Be that as it may, having overflow information at hand still does not unravel the issue. For ML models to provide sensible comes about, we not as it were got to bolster in expansive amounts of information but moreover ought to guarantee the quality of information. In spite of the fact that making sense out of crude information is an craftsmanship in itself and requires great include designing aptitudes and space information (in extraordinary cases), the quality information is of no utilize until it is appropriately utilized. The major issue which ML/DL professionals confront is how to partition the information for preparing and testing. In spite of the fact that it appears like a straightforward issue at to begin with, its complexity can be gaged as it were by plunging profound into it. Destitute preparing and testing sets can lead to unusual impacts on the yield of the demonstrate. It may lead to overfitting or underfitting of the information and our show may conclusion up giving one-sided comesabout

5.2.5. Applying Machine Learning Algorithm (CNN)

Here we used 2 convolution layers the 1st convolution layer contains 32 filters and the kernel size is 3 and the activation function we used here is the relu activation function and also we changed the input shape from 2D to 3D and the 2nd convolution layer which we used consists of 64 filters and the kernel size is 3 and also the activation function which is used here is relu. Then we used a pooling layer with max pooling with pool size equals to 2. Then we used a flatten layer It involves taking the pooled feature map that is generated in the pooling step and transforming it into a one-dimensional vector. This is done so that you can feed them as inputs to the dense layer. This output from the flatten layer becomes the input to the dense layer. The units in the dense layer 64 and the activation function used in the dense layer is relu activation function then this output from the dense layer becomes the input to the next dropout layer.

The Dropout layer is a mask that nullifies the contribution of some neurons towards the next layer and leaves unmodified all others. We can apply a Dropout layer to the input vector, in which case it nullifies some of its features; but we can also apply ito a hidden layer, in which case it nullifies some hidden neurons. Dropout layers are important in training CNNs because they prevent overfitting on the training data. If they aren’t present, the first batch of training samples influences the learning in a disproportionately high manner. This, in turn, would prevent the learning of features that appear only in later samples or batches: the dropout value which we took is 0.5.
5.2.6. Calculate Performance Parameters

This step consists of calculating different performance measures such as recall, precision, accuracy, etc., the model with the most acceptable is the one which has high accuracy and good recall and good precision is considered as a good model. The next step after implementing a machine learning algorithm is to find out how effective is the model based on the metric and datasets. Different performance metrics are used to evaluate different Machine Learning Algorithms. For example, a classifier used to distinguish between images of different objects; we can use classification performance metrics such as, Log-Loss, Average Accuracy, AUC, etc. If the machine learning model is trying to predict a stock price, then RMSE (root mean squared error) can be used to calculate the efficiency of the model. Another example of metric for evaluation of machine learning algorithms is precision recall or NDCG, which can be used for sorting algorithms primarily used by search engines. Therefore, we see that different metrics are required to measure the efficiency of different algorithms, also depending upon the dataset at hand.

MODEL VISUALIZATION

Visualizing a neural network can be helpful in understanding how different layers are connected to each other and how many neurons are there between each layer. It is one of the best ways to understand complex Neural Networks. Visualizing can also be helpful by presenting a model to someone to make them understand how different layers of a neural network are actually connected to each other. Some of the uses of visualization are:

1. Understanding how the model works.
2. Assistance in Hyper parameter tuning.
3. Finding out the failures of the model and getting an intuition of why they fail.
VI. RESULTS

CNN Results
After an extensive training process on a large dataset, the CNN model has achieved impressive results in terms of accuracy. The CNN model's superior performance in both training and testing phases validates its effectiveness as a powerful classifier, capable of accurately categorizing data into appropriate. The model's consistent and impressive results highlight its reliability and suitability for real-world scenarios, making it a promising choice for diverse machine learning and artificial intelligence applications.

Training Results
In the training phase after training the proposed model on the training dataset, these are the results which we have obtained. The highest training accuracy is observed at 45 epochs.

<table>
<thead>
<tr>
<th>Epochs</th>
<th>Training Accuracy</th>
<th>Training Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>99.92</td>
<td>0.37</td>
</tr>
<tr>
<td>10</td>
<td>99.89</td>
<td>0.43</td>
</tr>
<tr>
<td>15</td>
<td>99.92</td>
<td>0.45</td>
</tr>
<tr>
<td>20</td>
<td>99.91</td>
<td>0.37</td>
</tr>
</tbody>
</table>
Testing Results

These are the accuracies which we have obtained and the highest Testing accuracy is observed at 45 epochs.

<table>
<thead>
<tr>
<th>Epochs</th>
<th>Training Accuracy</th>
<th>Testing Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>99.92</td>
<td>99.91</td>
</tr>
<tr>
<td>10</td>
<td>99.89</td>
<td>99.90</td>
</tr>
</tbody>
</table>
VII Conclusion.

In this, we studied applications of machine learning like Convolution Neural Network and it shows that it proves accurate in deducting fraudulent transactions and minimizing the number of false alerts. If these algorithms are applied to a bank credit card fraud detection system, the probability of fraud transactions can be predicted soon after credit card transactions. And a series of anti-fraud strategies can be adopted to prevent banks from great losses and reduce risks. The objective of the study was taken differently than the typical classification problems in that we had a variable mis-classification cost. Precision, recall, f1-score, support, and accuracy are used to evaluate the performance of the proposed system. Comparing RandomForest Classifier with CNN and Logistic Regression, and we found out that the CNN is having giving more accuracy when compared to other algorithms Hence CNN is best Algorithm to Detect the Credit Card Frauds.

Acknowledgment

We want to express our deep-felt gratitude and sincere thanks to our guide Mrs.M.Madhavi, Assistant Professor, Department of AI, Anurag University, for her skilful guidance, timely suggestions, and encouragement in completing this project. We want to express our profound gratitude to all for having helped us in achieving this dissertation. Finally, we would like to express our heartfelt thanks to our parents, who were very financially and mentally supportive and for their encouragement to achieve our goals.

REFERENCES


