Topic: Credit Card Fraud Detection

Saiji Rabari

***Abstract*—Credit card extortion may be a predominant issue in monetary exchanges, posturing noteworthy challenges to both budgetary educate and customers. With the rise of computerized installments, conventional strategies of extortion location have gotten to be insufficient, requiring the investigation of progressed methods such as machine learning. In this inquire about, we explore the viability of different machine learning models in identifying credit card extortion employing a freely accessible dataset.We employ logistic regression, shallow neural networks outfit strategies like random forests and gradient boosting, and support vector machines to classify fraud and non-fraudulent exchanges. Our discoveries uncover the qualities and impediments of each show, giving bits of knowledge into the ideal approach for credit card extortion discovery.**

Keywords — Machine Learning(ML), Deep Learning(DL), Shallow Neural Netowrks, Logistic Regression, Random Forests, Gradient Boosting, Support Vector Machines(SVM)

I. Introduction

- **Requirement of Fraud Detection** : With the expanding selection of computerized installment strategies, counting credit cards, occurrences of fraud have been on the rise. Concurring to the Reserve Bank of India (RBI), the number of credit card fraud cases detailed by banks expanded from 6,801 cases in 2018-19 to 14,498 cases in 2019-20

  Credit card fraud leads to noteworthy money related misfortunes for both people and banks. In 2019-20, the sum misplaced due to credit card fraud in India summed to 133.8 crore, as detailed by the RBI. These misfortunes can have a serious affect on individuals' money related well-being and can disintegrate believe within the managing an account framework.

  In today's computerized age, the utilize of credit cards has ended up omnipresent, advertising comfort and adaptability in budgetary exchanges. Be that as it may, this comfort too comes with dangers, as credit card extortion remains a predominant issue influencing both shoppers and budgetary institutions.

- **Range :** Credit card fraud includes unauthorized exchanges made utilizing stolen or fake credit card data, coming about in money related misfortunes and potential hurt to consumers' credit scores. Agreeing to industry reports, billions of dollars are misplaced every year to fraudulent activities, making it a noteworthy concern for stakeholder within the budgetary sector.

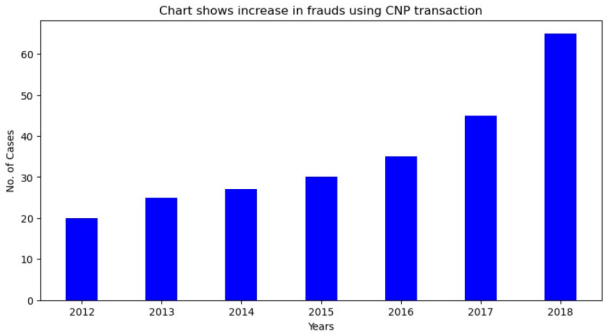

Figure 1: Increased Number of cases for fraud of Credit Card transactions over the year

- **Challenges Faced:** Ordinary procedures of Extortion discovery, such as rule-based systems and manual overviews, are routinely inadequately in recognizing cutting edge extortion works out. Fraudsters tirelessly progress their techniques, abusing vulnerabilities in existing discovery frameworks and altering to present day security measures.

- **Impact's:** In later a long time, there has been a developing intrigued in leveraging machine learning methods to address the challenges of credit card extortion detection. Machine learning calculations offer the potential to analyze huge volumes of value-based information, recognize designs characteristic of false behavior, and adjust to advancing extortion plans in real-time.

- **The objective of this research:** is about to explore the adequacy of different machine learning models in recognizing credit card extortion. By assessing and comparing diverse models, we point to distinguish the foremost reasonable approach for moderating the dangers related with false exchanges.

II. Methodology:-

- **Retrieval of Data & Processing of it:-** We gotten the credit card exchange dataset from a trustworthy source, containing highlights such as exchange amount,time.

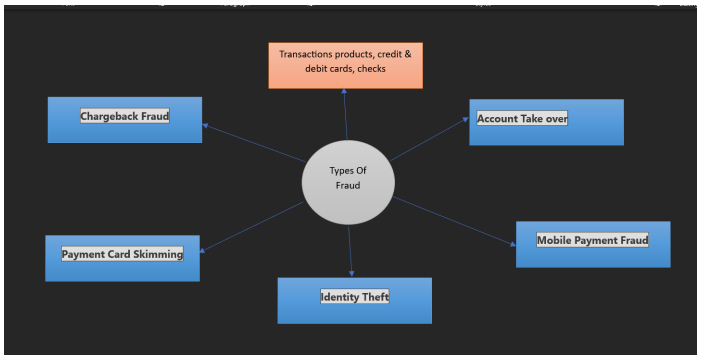

Figure 2: Types Of Fraud

We performed pre-processing steps counting information cleaning, taking care of misplaced values, and include scaling to arrange the dataset for demonstrate training.

- **Pedagogy of Analysis and Result:-** We chose a assortment of machine learning models for our research, counting calculated relapse, shallow neural networks, arbitrary woodlands, angle boosting, and back vector machines. Each show was chosen based on its suitability for double classification assignments and past victory in fraud detection applications
  We portion the dataset into planning, testing, and endorsement sets employing a stratified approach to ensure balanced less on dissemination. We arranged each illustrate on the preparing information and evaluated its execution utilizing fitting measurements such as accuracy, precision, survey, and F1-score
- **An Outcome from the Analysis:-** We analyzed the results gotten from each show to evaluate its practicality in distinguishing credit card extortion. We compared the execution measurements of unmistakable models and talked about their qualities and deficiencies. Bits of information picked up from this examination were utilized to recognize the foremost promising approach for credit card extortion location.

III. Overview of Existing Research:-

Early research in credit card fraud detection predominantly centered on rule-based frameworks and manual audit forms. These frameworks depended on predefined rules and heuristics to hail suspicious exchanges for assist examination. Whereas compelling to a few degree, they regularly battled to keep pace with the energetic nature of fraudulent activities

- **Challenges:** In spite of the guarantee of machine learning, a few challenges exist in conveying these models for real-world fraud discovery. One major challenge is the awkwardness between fraudulent and legitimate transactions, driving to one-sided models that favor the lion's share lesson. Also, guaranteeing demonstrate interpretability and straightforwardness is vital for picking up stakeholders' believe and compliance with administrative prerequisites.
- **Recent research:** has centered on tending to these challenges through procedures such as class imbalance correction, include building, and show explainability. Gathering strategies like random forests and gradient boosting have picked up notoriety for their capacity to handle imbalanced datasets and create precise forecasts.

IV. Overview on our approach

- **Comprehensive Assessment:-** Our research gives a comprehensive assessment of different machine learning models for credit card fraud detection. By comparing the execution of logistic regression, shallow neural networks, random forests, gradient boosting, and support vector machines, we offer bits of knowledge into the qualities and shortcomings of each approach.
  Not at all like simply hypothetical ponders, we illustrate the practical usage of machine learning models utilizing real-world credit card transaction data. The code given in our examination exhibits how these models can be prepared, assessed, and sent in a generation environment.
- **Code-Based Solution:-**By analyzing the given code, we offer a commonsense viewpoint on the usage and execution of machine learning models for credit card fraud detection. This hands-on approach permits perusers to get it the complexities included in model advancement and pick up experiences into best practices for handling real-world challenges.
- **Model Comparison:-**Through our examination, we compare the adequacy of diverse machine learning models in recognizing credit card fraud. By looking at measurements such as precision, accuracy, recall, and F1-score, we distinguish the foremost appropriate model for fraud detection, considering variables such as interpretability, versatility, and computational productivity.

V. Results & Analysis

A. System Specifications:-

- **Processor :-** Intel® Core™ i5-12450H, 8C (4P + 4E) / 12T, P-core 2.0 / 4.4GHz, E-core 1.5 / 3.3GHz, 12MB
- **Graphics:-** NVIDIA® GeForce RTX™ 3050 4GB GDDR6, Boost Clock 1500MHz, TGP 85W
- **Memory:-** 2x 8GB SO-DIMM DDR4-3200

B. Description Of Our Project

We have used already exist model like logistic model, Linear Support Vector Machine(SVM) model, Gradient Boosting Classifier and Random Forest model.We Employed Shallow neural network model with value of relu=2. At layer 2, output shape will be turned to two nodes , at dense layer-3 it will turned to one node.

To determine which model is more proficient have to compare them by parameters such as accuracy, perfomance, F1-score and precision.

The confusion matrix in Figure summarizes the performance of the classification model in distinguishing Fraud transactions from legitimate ones (Not Fraud). Each cell in the matrix represents a combination of predicted and actual classes:-

- True Positive (TP): The number of correctly classified instances of fraud. These are transactions that were correctly identified as fraudulent.

- False Negative (FN): The number of instances of fraud that were incorrectly classified as not fraud. These are fraudulent transactions that went undetected by the model.
- False Positive (FP): The number of instances incorrectly classified as fraud when they are not. These are legitimate transactions that were flagged as fraudulent, potentially causing inconvenience to customers.
- True Negative (TN): The number of correctly classified instances of not fraud. These are legitimate transactions correctly identified as such.

Understanding these values is vital for assessing the adequacy of the demonstrate in recognizing fraudulent transactions whereas minimizing false positives and false negatives, hence guaranteeing the security and dependability of the money related framework.

		Predicted Class	
		Predicted Fraud	Predicted Not Fraud
Actual Class	Fraud (+)	True Positive (TP)	False Negative (FN)
	Not Fraud (-)	False Positive (FP)	True Negative (TN)

Figure 3: Confusion Matrix

	precision	recall	f1-score	support
Not Fraud	1.00	1.00	1.00	22771
Fraud	0.73	0.53	0.61	36
accuracy			1.00	22807
macro avg	0.87	0.76	0.81	22807
weighted avg	1.00	1.00	1.00	22807

Figure 4: Result Obtained From logistic model

C. Analysis On Values Of Confusion Matrix:-

If there is true indication for not fraud than there should be false for all field of fraud.In output of shell number 13 in our code, we got 0.73 precision which indicates false positive. In real world, If transaction is not fraud & our model predict it is fraud than it is considered as 'misclassifiaction'. So Confusion matrix and Accuracy are not enough for better solution, with them we have to also consider precision, recall & F1-score.

D. Employed Shallow neural network:-

```
shallow_nn_b = Sequential()
shallow_nn_b.add(InputLayer((x_train.shape[1],)))
shallow_nn_b.add(Dense(2, 'relu'))
shallow_nn_b.add(BatchNormalization())
shallow_nn_b.add(Dense(1, 'sigmoid'))

checkpoint = ModelCheckpoint('shallow_nn_b',
                             save_best_only=True)
shallow_nn_b.compile(optimizer='adam',
         loss='binary_crossentropy',  metrics=['accuracy'])
shallow_nn_b.fit(x_train_b, y_train_b,
   validation_data=(x_val_b, y_val_b),epochs=40, callbacks=checkpoint
```

Figure 5: Creation of Shallow neural network

As we see earlier logistic model is not much suitable for fraud detection. so we have employed shallow neural network model.Above Figure shows, code written by us to create this model.This model will give us desirable results.

E. Prediction Function:-

As We mention in above section, we can't use accuracy as predictable measurement so will make prediction function that make probability into the class. In Figure 4, We have described prediction function that we use for creation of shallow neural network.

```
#17
def neural_net_predictions(model, x):
    # Make predictions using the neural network model on the input
    predictions = model.predict(x)

    # Flatten the predictions array and compare each element with 0
    # If prediction > 0.5, classify as 1 (Fraud), otherwise classif
    binary_predictions = (predictions.flatten() > 0.5).astype(int)

    return binary_predictions

# Call the neural_net_predictions function with the shallow neural
predictions = neural_net_predictions(shallow_nn, x_val)

# Print the predictions
print(predictions)
```

Figure 6: Prediction Function

F. Random forests:-

Random forests are an outfit learning strategy that works by developing a large number of decision trees amid preparing and yielding the lesson that's the mode of the classes (classification) of the individual trees. They are vigorous against overfitting and are competent of taking care of high-dimensional datasets with categorical and numerical highlights.

Class	Precision	Recall	F1-Score
Not Fraud	1.00	1.00	1.00
Fraud	0.81	0.47	0.60
Accuracy			1.00
Macro Avg	0.90	0.74	0.80
Weighted Avg	1.00	1.00	1.00

Table I: Performance Metrics for Random forests

G. Gradient Boosting Classifier:-

Class	Precision	Recall	F1-Score	Support
Not Fraud	1.00	1.00	1.00	22771
Fraud	0.67	0.67	0.67	36
Accuracy			1.00	22807
Macro Avg	0.83	0.83	0.83	22807
Weighted Avg	1.00	1.00	1.00	22807

Table II: Performance metrics for Gradient boosting

Gradient boosting is another outfit learning procedure that builds a solid prescient model by combining the forecasts of different frail models, ordinarily decision trees. It iteratively moves forward the demonstrate by minimizing a loss function, centering on occasions that were already misclassified. Gradient boosting is known for its high predictive accuracy and robustness.
In Gradient boosting, Precision is high which is good and if less recall is not problem than one can go for gradient boosting.

H. Linear Support Vector Machine:-

Support Vector Machines are capable administered learning models utilized for classification and regression tasks. They work by finding the hyperplane that best isolates the classes within the highlight space, maximizing the edge between the classes. SVMs are compelling in high-dimensional spaces and are especially valuable when the number of features surpasses the number of samples

As we can see from below table this model has good combination of precision, recall, F1-score & accuracy.

	Precision	Recall	F1-score
Not Fraud	1.00	1.00	1.00
Fraud	0.65	0.78	0.71
Accuracy			1.00
Macro avg	0.83	0.89	0.85
Weighted avg	1.00	1.00	1.00

Table III: Performance metrics for Linear SVM model

VI. COMPARISON:-

Before Comparing performance metrics, we have to transfer imbalanced data set into balanced data set.After that we applied Shallow_nn, Random forest, Gradient Boosting Classification and Linear SVC on this balanced data set.

A. Significance of Precision:-

High Precision means model not predict fraud when it is not fraud.
If precision is low then model will predict fraud although it is not fraud.
If Recall is less then model can't identify fraud when it is originally fraud.
Models except Shallow_nn has less recall. Shallow_nn has 0.87 recall on balanced data set.
Shallow_nn has precision value 0.94 and other models have 1.00 which is good but there is difference of only 0.06 which can be ignored because recall value of other model is very less.

Table IV: Performance Metrics for Different Models on balanced data set

Model	Metric	Not Fraud	Fraud
Logistic Regression	Precision	0.84	0.96
	Recall	0.95	0.84
	F1-Score	0.89	0.90
Shallow Neural Network	Precision	0.86	0.94
	Recall	0.94	0.87
	F1-Score	0.90	0.90
Random Forest	Precision	0.65	1.00
	Recall	1.00	0.53
	F1-Score	0.79	0.69
Gradient Boosting	Precision	0.66	1.00
	Recall	1.00	0.55
	F1-Score	0.80	0.71
Support Vector Machine	Precision	0.79	1.00
	Recall	1.00	0.76
	F1-Score	0.88	0.87

B. Conclusion:-

Shallow_nn has good combination of precision, recall & F1-score.so we will go for shallow_nn with relu = 2.

- The framework tracks online exchanges and recognizes clients based on their client ID and password.It shows up that the framework moreover stores data approximately the client such as title, address, versatile number, and email address which may be utilized to confirm personality amid fraud detection.
- Admins can see exchanges and possibly oversee clients based on the data within the model.Overall, this ER show chart recommends a rearranged framework for online exchange fraud transaction. It centers on client verification and exchange following, which are basic components of numerous fraud detection systems.
- Here are a few extra subtle elements to consider. This data would be significant for distinguishing fraud transactions.The show moreover needs any entities related to fraud detection strategies or rules. It's conceivable that this data is put away somewhere else within the system.Without a more comprehensive see of the framework, it's troublesome to say authoritatively how fraud detection is taken care of. In any case, the ER demonstrate gives a beginning point for understanding the center components.

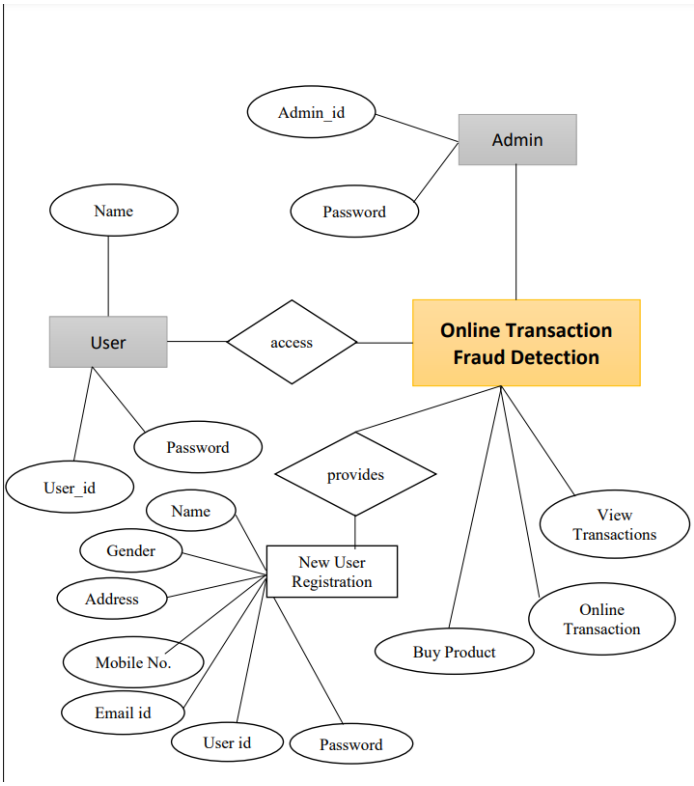

Figure 7: ”Unlocking insights: ER model delivers clarity and structure to data like never before!”