Credit card fraud detection Using Machine Learning Algorithms

Abstract: We are continuously moving toward the online banking system. All the things were converted into an online mode where users can do transactions anywhere at any time. But the rate of cyber-crime and fraud is increasing day by day. Our project's main purpose is to make people aware of the ongoing online credit card fraud. The main point of the credit card fraud detection system is the necessity to safeguard our transactions and improve security. With the solution provided by our research, the fraudsters won't have a chance to make multiple transactions from a stolen or counterfeit card. Instead, the cardholder will get aware of the fraud before any activity by the fraudster. The model can also detect whether any new transaction is fraudulent or not. Our aim here will be to provide a model to detect 100% of the fraudulent transactions while minimizing the incorrect fraud classifications.

Introduction:

A credit card is a card which is work as an instrument through which users can do online transactions. It is provided by a financial institution or by an organization, it allows users to borrow funds. The limit of the credit card is determined by the credit score, income, and credit history of the user. It can be used for shopping, electricity bills, restaurants, electronic devices, etc.

Credit Card Fraud:

It refers to a scammer using your card number and pin for transactions without your knowledge, or they have stolen the card for financial transactions from your account.

Credit card fraud comes in many different shapes and forms, including fraud that involves using a payment card of some description, and more. The reasons for credit card fraud also vary. Some are designed to obtain funds from accounts, while others wish to obtain goods for free. Furthermore, it is very important to understand that credit card fraud is linked closely to identity theft. According to the Federal Trade Commission, some 5% of all people over 16 in this country have been or will be the victim of identity theft. Additionally, at the last count in 2008, it was found that there had been a 21% growth in the prevalence of identity theft. On the other hand, the percentage of identity theft cases related to credit card fraud decreased, which is a positive thing and a credit to law enforcement professionals and the general public as a whole.

Types of credit card fraud:

1. Application Fraud

Application fraud generally happens in conjunction with identity theft. It happens when other people apply for credit or a new credit card in your name. They will usually first steal supporting documents, which are then used to substantiate their fraudulent application. Banks have various safeguarding measures in place to stop this type of fraud from happening. The most important one is requiring original documentation only. Additionally, they will often telephone employers to confirm identity. Unfortunately, criminals will frequently forge documents and provide false telephone numbers for places of employment. Unfortunately, there are always ways around certain safeguarding measures.

2. Electronic or Manual Credit Card Imprints

A second form of credit card fraud is experienced through credit card imprints This means that somebody skims information that is placed on the magnetic strip of the card. This is then used to encode a fake card or to complete fraudulent transactions.

3. CNP (Card Not Present) Fraud

If somebody knows the expiry date and account number of your card, they can commit CNP fraud against you. This can be done through phone, mail or internet. It essentially means that somebody uses your card without actually being in physical possession of it. More and more and often,

merchants will require the card verification code, making CNP fraud slightly more difficult, but if a fraudster can get your account number, they probably know that number too. Additionally, there are only 999 possible combinations for the verification code. As such, many criminals attempt to order items of very low amounts until they figure out the right number. Be on the lookout, therefore, for small payments on your statements.

4. Counterfeit Card Fraud

Counterfeit card fraud is usually committed through skimming. This means that a fake magnetic swipe card holds all your card details. This fake strip is then used to create a fraudulent card that is fully functional. Essentially, it is an exact copy, which means fraudsters can simply swipe it in a machine to pay for certain goods. This type of fraud can also be committed by someone who knows your card details. They can use this information to create a so-called 'fake plastic'. Here, the magnetic strip or the chip on the card doesn't actually work. However, it is often easy enough to convince a merchant that there is something wrong with the card, at which point they will enter the transaction by hand.

5. Lost and Stolen Card Fraud

The next possible type is lost and stolen card fraud. Here, your card will be taken from your possession, either through theft or because you lost it. The criminals who get their hands on it will then use it to make payments. It is difficult to do this through machines, as they will require a pin number. However, it is easy enough to use a found or stolen card to make online purchases. It is for this reason that it is vital that you cancel your cards as soon as you realize they are missing.

6. Card ID Theft

Card ID theft happens when the details of your card become known to a criminal, and this information is then used to take over a card account or open a new one. Your name will be used for this. This is one of the most difficult types of fraud to identify and to recover from, because it can take a long time before you even know that it has happened.

7. Mail Non-Receipt Card Fraud

This type of fraud is also known as never received issue or intercept fraud. In this case, you were expecting a new card or replacement one and a criminal is able to intercept these. The criminal will then register the card and they will use it to make purchases and more.

8. Assumed Identity

With assumed identity fraud, a criminal will use a temporary address and a false name to obtain a credit card. There are a number of systems in place with banks for protection against this type of fraud. For instance, they will ask new customers to provide account references and these will be check to ascertain that they are genuine. Additionally, they could ask for such things as birth certificates, original copies of driver's license or passports and so on. They often ask for these things before they will send a card out.

9. Doctored Cards

A doctored card is a card whereby a strong magnet has erased its metallic stripe. Criminals do this and then manage to change the details on the card itself so that they match those of valid cards. Naturally, this card won't work when a criminal tries to pay for something. However, they will then use their charm to convince a merchant to just enter the details of the card manually.

10. Fake Cards

It takes a lot of time, skill and effort to create fake credit cards, but that doesn't stop a determine criminal. A card has to meet certain complex security features and cards are becoming increasingly advanced, meaning this is much harder to do. There is the magnetic strip, the chip and, often, holograms. However, someone who is skilled can forge this type of cards using fake names and numbers and will make transactions with the card. The card isn't actually linked to an account, so the credit card company will not pay for the transaction since they cannot link it to a specific user. By that

time, however, the criminal will be long gone with their purchases.

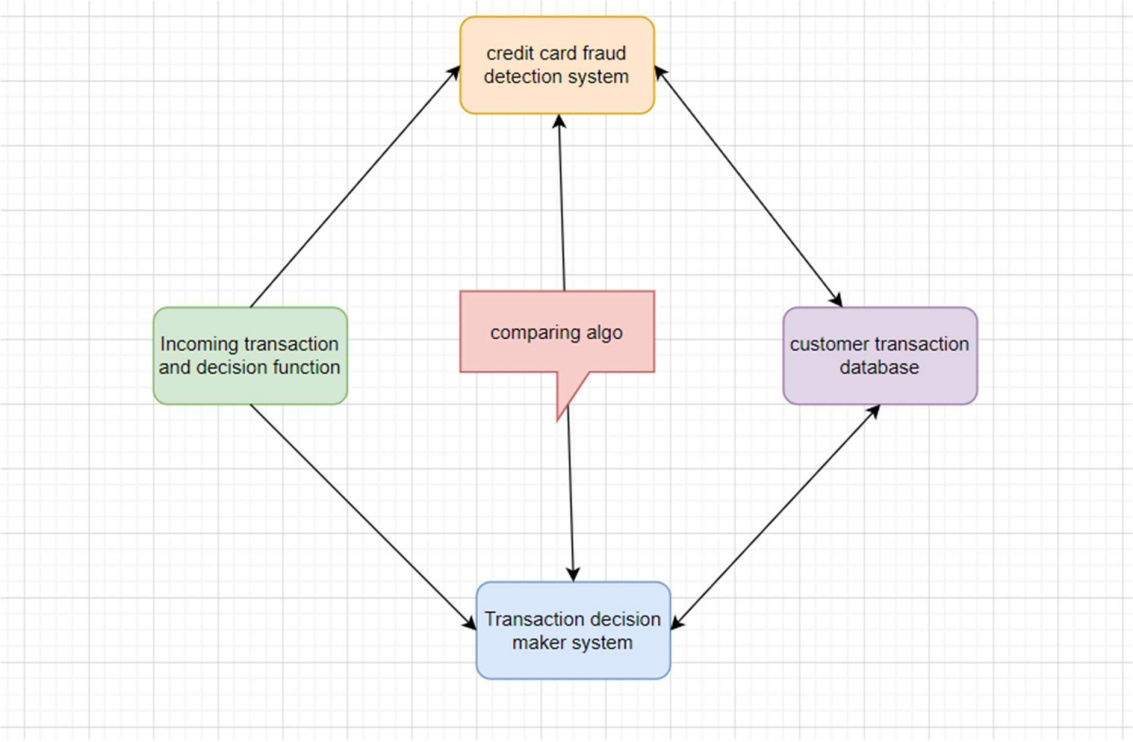
11. Account Takeover

Account takeover is actually one of the most common forms of credit card fraud. Basically, a criminal will somehow manage to get hold of all of your information and relevant documents. This is usually done online. They will then contact the credit card company and pretend to be you, asking them to change the address. They will provide 'proof' of identity, since they have hacked through or otherwise obtained, your personal details. A replacement card will then be sent to the fake address, and the criminal will be able to make charges.

Unfortunately, it isn't rare for this type of fraud to occur. It is important, therefore, that you are aware of what they are and you must be able to take the appropriate steps to prevent criminals from committing credit card fraud against you. Protecting your personal information is the most important element of that. This means common sense steps such as using strong, unique passwords and not leaving documents in plain sight.

OBJECTIVES: The objective of the project is to implement machine learning algorithms to detect credit card fraud detection with respect to time and amount of transaction.

Proposed solution: The basic rough architecture diagram can be represented with the following figure:



Fig[1]: Architecture Diagram for credit card fraud detection

Machine learning algorithms used in detection of fraud:

1 Logistic Regression: Logistic regression works with sigmoid function because the sigmoid function can be used to classify the output that is dependent feature and it uses the probability for classification of the dependent feature. This algorithm works well with less amount of data set because of the use of sigmoid function if value the of sigmoid function is greater than 0.5 the output will 1 if the output the sigmoid function is less than 0.5 then the output is considered as the 0. But this sigmoid function is not suitable for deep learning because the if deep learning when we back tracking from the output to input we have to update the weights to minimize the error in weight update. we have to do differentiation of sigmoid activation function in middle layer neuron then results in the value of 0.25 this will affect the accuracy of the module in deep learning.

2 Decision Tree:Decision tree can be used for the classification and regression problems working for both is same but some formulas will change. Classification problem uses the entropy and information gain for the building of the decision tree model. entropy tell about how the data is random and information gain tells about how much information we can get from this feature.Regression problem uses the gini and gini index for the building of the decision tree model. In classification problems the root node is selected by using information gain that the root node t id selected by using is having the high information again and low entropy. In Regression problems the root node is selected by using gini , the feature which is having the less gini is selected as the root here Depth of the tree can be determined by using hyper parameter optimization, this can be achieved by Using grid search cv algorithm.

3 Random Forest: The random forest randomly selects the features that is independent variables and also randomly selects the rows by row sampling and the number of decision tree can be determined by using hyper parameter optimization. For classification problem statement the output is the maximum occurrence outputs from each decision tree models inside the random forest. This is one the widely used machine learning algorithm in real word scenarios and in deployed models. And in most of the Kaggle computation challenges this algorithm is used to solve the problem statement.

4 Naive Bayes: Naive Bayes is the machine learning algorithm for classification problem, which work on the property of Bayes theorem. It can be implemented by using features in data set independent feature as input and dependent feature as a output, the same thing what is behind the Naive Bayes theorem is applied here to calculate probability of the dependent feature with respect to independent features.

Result:

Importing the Libraries

Several libraries will be used for different purposes in this post. The pandas library will be used to load the data into
a DataFrame object making it easier to work with. The matplotlib and seaborn libraries will be used for plotting purposes. While
the sklearn library will be used to perform some data processing, model building, and model evaluation.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, precision_score, recall_score
from sklearn.metrics import confusion_matrix
```

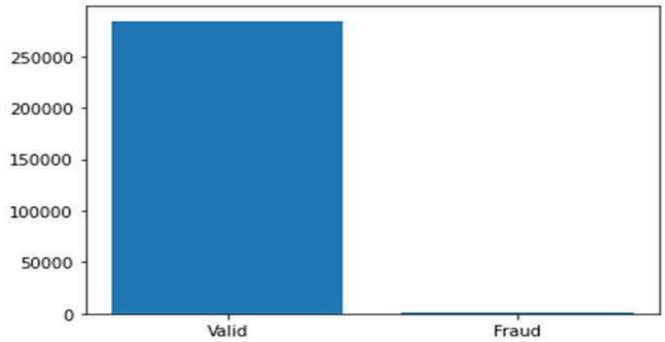
Performing Exploratory Data Analysis

The dataset that will be used for credit card fraud detection using machine learning is available here: Credit Card Fraud Detection Data. The dataset consists of 30 features – time, predictors V1 to V28, and amount. The columns V1 to V28 likely consist of sensitive credit card information and hence have been anonymized and scaled. The class column is the column to be predicted where 0 represents a valid transaction and 1 represents a fraudulent one.

Before using the data to train the machine learning model, it is better to understand the data we are dealing with. This step is known as an exploratory data analysis and usually includes steps like determining the shape of the dataset i.e. the number of rows and columns, identifying the type of data objects in each column, identifying the missing values, determining the correlation values, etc.

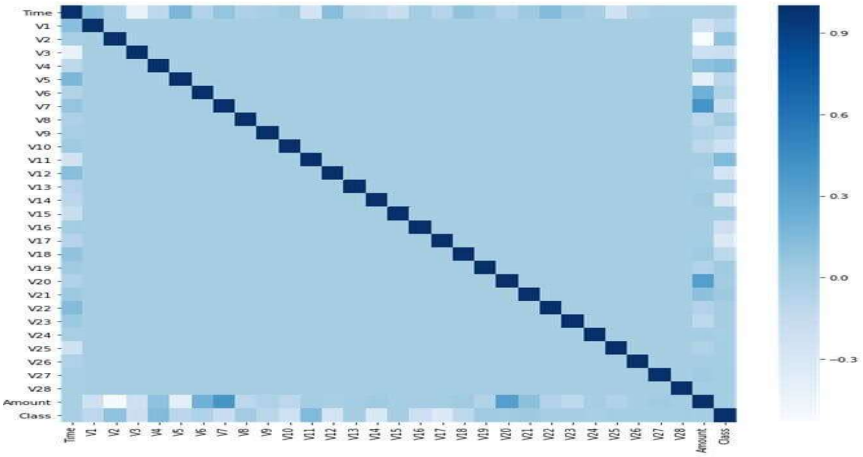
```
data=pd.read_csv('creditcard.csv')
plt.bar(['Valid','Fraud'],list(data['Class'].value_counts()))
print("Fraudulent transactions: ", end='')
frauds= data['Class'].value_counts()[1]/sum(data['Class'].value_counts())
print(round(frauds*100,2), end='%')
plt.show()
```

The dataset appears to be highly unbalanced with fraudulent transactions only representing 0.17% of all transactions. Unbalanced datasets may lead to bias in machine learning models and hence should be handled accordingly.



Determining the extent of correlations between the variables in our dataset is very useful information. This information can help with which features to extract or which machine learning model to select. Plotting the correlation matrix provides a visual summary of the correlation values between the features and the outcome.

```
fig=plt.figure(figsize= (12, 12))
sns.heatmap(data.corr(), cmap='Blues')
plt.show()
```



The heat-map above indicates that there are no high correlation values among the predictor columns. No predictor column has a high correlation value with the Class column either. However there exists a negative correlation among V2 and Amount as well as a positive correlation among V7 and the Amount feature.

Data Pre-Processing

Data pre-processing involves preparing the dataset to train the machine learning model. The data pre-processing step is crucial and should transform the data in a way that can be processed by the selected machine learning algorithm. For example, most classification algorithms will not be able to understand the text in the data, and hence not performing data pre-processing will lead to errors.

Common data pre-processing steps involve – imputing or dropping records containing missing values, label encoding categorical data, one hot encoding labeled data, scaling the data, and performing train-test splits on the dataset.

As this dataset does not contain any missing values or categorical data, most data pre-processing steps are not needed. The data is taken and first split into the predictors i.e. X and the outcome i.e. Y. X contains 284807 data records with 30 features each while Y contains 284807 data records with one column – class.

```
X =data.iloc[:,:-1]
Y =data.iloc[:,-1]
X_train, X_test, Y_train, Y_test=train_test_split(X, Y, test_size=0.2,
random_state=42)
```

The train-test split divides the dataset into a training set and testing set. The training set is used to train the machine learning model while the testing set is used to evaluate the model. The test size of 0.2 indicates that 20% of the dataset is chosen to be the testing set. Hence, the training set contains 227845 records while the testing set contains 56962 records.

Classification Model

Selecting a machine learning model depends on the type of task that is required to be performed. Machine learning can perform various tasks such as classification, regression, clustering, pattern extraction, etc. Within each task there are a number of algorithms that may be available. Usually, two or more algorithms are experimented with to decide which model suits the data better and gives more accurate and robust results.

The credit card fraud detection problem is a classification problem, as it involves classifying a credit card transaction to be in either of the two classes – valid or fraudulent. As mentioned, there are several classification algorithms available such as Linear Classifiers, Naïve Bayes Classifier, Support Vector Machines, Nearest Neighbour Classifier, Decision Trees, etc. For this problem, a Random Forest Classifier is implemented which is an extension of the Decision Tree classifier.

```
classifier=RandomForestClassifier()
classifier.fit(X_train, Y_train)
Y_pred=classifier.predict(X_test)
```

In the above code, an object of the RandomForestClassifier class belonging to the sklearnlibrary is created. Using the *fit* function of this class, the model is trained using the training set. Finally, the *predict* function gives a prediction for the values of features in the testing set.

Model Evaluation

Every machine learning model must be evaluated on the task that it performs. Model evaluation involves asking the model to predict the values for unseen data records based on what it has learnt. This has been done above and is stored in Y_pred. Y_pred are the values as predicted by the model which must be compared against the true values i.e. Y_test. As this is a classification problem, we can evaluate the model using metrics such as accuracy, precision, and recall.

```
print("Model Accuracy:", round(accuracy_score(Y_test, Y_pred),4))
print("Model Precision:", round(precision_score(Y_test, Y_pred),4))
print("Model Recall:", round(recall_score(Y_test, Y_pred),4))
```

An accuracy of the model determines how many data records the model predicted correct values. This model has an accuracy value of 0.9996 which indicates that the model made

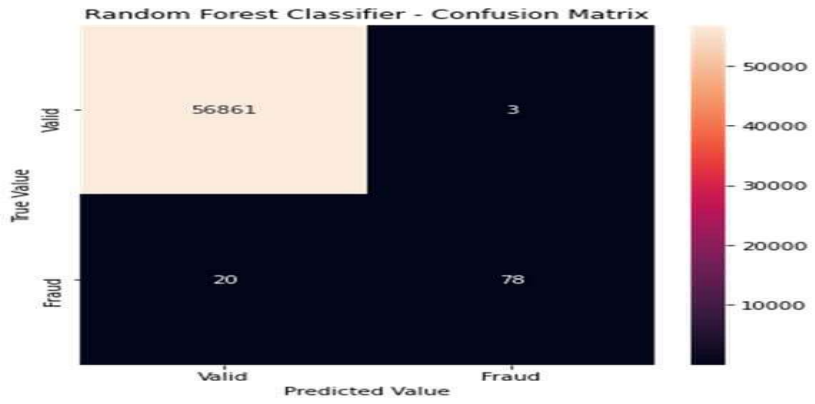
correct predictions 99.96% of the time. Precision indicates the correctness of all those records that were predicted to be positive. A precision value of 0.963 indicates that when the model predicted a positive result it was correct 96.3% of the time. Lastly, recall of a model indicates how many truly positive values were identified correctly. A recall value of 0.7959 indicates that the model was able to identify 79.59% of all positive values correctly.

The metrics discussed above can be visualized using a heat-map on something known as the confusion matrix. A confusion matrix displays the values of the number of predictions between the true and predicted values of each class.

```
labels= ['Valid', 'Fraud']
conf_matrix=confusion_matrix(Y_test, Y_pred)
plt.figure(figsize=(6, 6))
sns.heatmap(conf_matrix, xticklabels= labels, yticklabels= labels, annot=True,
fmt="d")
plt.title("Random Forest Classifier - Confusion Matrix")
plt.ylabel('True Value')
plt.xlabel('Predicted Value')
plt.show()
```

As seen from the confusion matrix, the model was correctly able to classify 56861 records as valid and 78 records as fraudulent. However, it incorrectly identified a valid transaction as a fraudulent transaction 3 times and incorrectly identified a fraudulent transaction as a valid transaction 20 times

.



Conclusion:

our machine learning classifier was able to classify the validity of credit card transactions with a 99.6% accuracy.

References:

1. S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang, and G. N. Surname, Random forest for credit card fraud detection, IEEE 15th International Conference on Networking, Sensing and Control (ICNSC),2018.
2. Satvik Vats, Surya Kant Dubey, Naveen Kumar Pandey, A Tool for Effective Detection of Fraud in Credit Card System, published in International Journal of Communication Network Security ISSN: 2231 1882, Volume-2, Issue-1, 2013.
3. Rinky D. Patel and Dheeraj Kumar Singh, Credit Card Fraud Detection & Prevention of Fraud Using Genetic Algorithm, published by International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-6, January 2013.
4. M. Hamdi Ozcelik, Ekrem Duman, Mine Isik, Tugba Cevik, Improving a credit card fraud detection system using genetic algorithm, published by International conference on Networking and information technology, 2010.
5. Wen-Fang YU, Na Wang, Research on Credit Card Fraud Detection Model Based on Distance Sum, published by IEEE International Joint Conference on Artificial Intelligence, 2009.
6. Andreas L. Prodromidis and Salvatore J. Stolfo; "Agent-Based Distributed Learning Applied to Fraud Detection"; Department of Computer Science- Columbia University; 2000.
7. Salvatore J. Stolfo, Wei Fan, Wenke Lee and Andreas L. Prodromidis; "Cost-based Modeling for Fraud and Intrusion Detection: Results from the JAM Project"; 0-7695-0490-6/99, 1999 IEEE.
8. Soltani, N., Akbari, M.K., SargolzaeiJavan, M., A new user-based model for credit card fraud detection based on artificial immune system, Artificial Intelligence and Signal Processing (AISP), 2012 16th CSI International Symposium on., IEEE, pp. 029-033, 2012.
9. S. Ghosh and D. L. Reilly, Credit card fraud detection with a neural- network, Proceedings of the 27th Annual Conference on System Science, Volume 3: Information Systems: DSS/ Knowledge Based Systems, pages 621-630, 1994. IEEE Computer Society Press.
10. MasoumehZareapoor, Seeja.K.R, M.Afshar.Alam, Analysis of Credit Card Fraud Detection Techniques: based on Certain Design Criteria, International Journal of Computer Applications (0975 8887) Volume 52 No.3, 2012.
11. Fraud Brief AVS and CVM, Clear Commerce Corporation, 2003, http://www.clearcommerce.com.
12. All points protection: One sure strategy to control fraud, Fair Isaac, http://www.fairisaac.com, 2007. [13] Clear Commerce fraud prevention guide, Clear Commerce Corporation, 2002, http://www.clearcommerce.com
13. Samaneh Sorournejad, Zahra Zojaji , Reza Ebrahimi Atani , Amir Hassan Monadjemi, A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective , IEEE 2016
14. S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang, and G. N. Surname, Random forest for credit card fraud detection, IEEE 15th International Conference on Networking, Sensing and Control (ICNSC),2018.