Credit Card Fraud Identification

Riya Dubey, Siddhant Jain, Parth Sharma, Ms. Bhaivee Singh

Submitted: 10-05-2022 Revised: 15-05-2022 Accepted: 18-05-2022

ABSTRACT—With the increasing technology, the use of smart cards like credit cards has become popular. In addition, credit card abuse and fraud have also come to the fore. Such fraudulent use may harm the user. Know that his / her credit card is completely secure. The purpose of our project is to detect credit card fraud. Any malicious person using our card can be detected and thus fraudulent use detected.

Index Terms—cheating, KNN algorithm, logistic regression algorithm, SVM algorithm, random forest algorithm, ROC curve, accuracy.

I. INTRODUCTION

PHYTON is a widely used, well-defined, general-purpose, high-level programming language developed by Guido van Rossum in 1991. More popular. And widely used language. It takes a few lines to implement compared to other programming languages. It is basically. Python is a programming language that works quickly and integrates systems efficiently. Python is used to develop web applications as well as complex scientific applications. Python can be used to analyse data and display a large number of libraries to visualize data through those libraries.

In computer science, artificial intelligence (AI) is sometimes defined as machine language, which refers to providing intelligence to a machine that mimics human behaviour. According to John McCarthy, the father of artificial intelligence, it was "the science and engineering that made intelligent machines, especially intelligent computer programs."

Credit card fraud is a term used to describe fraud committed using a credit card, such as a payment card. A credit card is a payment card issued to a customer that allows the cardholder to pay the merchant for goods and services based on a promise that the card issuer will pay the promised amount and other agreed charges.

The issuer creates a revolving account and presents a line of credit to the cardholder, allowing the cardholder to borrow from the merchant for payment or as a cash advance. 'Fraud' is the unauthorized and unnecessary use of someone else's account in a credit card transaction. Than that account owner. Necessary preventive measures can be taken to prevent this abuse and it can be minimized by studying the behaviour of such fraudulent practices and preventing similar incidents from happening in the future. In other words, credit card fraud can be defined as the use of another person's credit card for personal reasons, but the owner and the issuing authority of the card are unaware that the card is being used. Fraud detection involves monitoring activities. Predict, understand or prevent consumer abuse, including fraud, intrusion and negligence. There are two types of card fraud - card-present fraud and card-present fraud. Compromise comes in many forms and usually without the knowledge of the cardholder. Cardholders can quickly report stolen cards, but fraudulent account details can remain with the fraudster for months before any theft, making it difficult to trace the source of the compromise. Fraudulent usage may not be detected until the cardholder receives the statement. Cardholders can reduce the risk of this fraud by frequently checking their accounts to make sure there are no suspicious or unknown transactions. When a credit card is lost or stolen, the holder notifies the issuing bank and uses it for illegal purchases unless the bank account is blocked.

II. LITERATURE REVIEW-

Data mining technology is a great way to solve the problem of credit fraud detection. Credit card fraud detection is the process of identifying fraudulent transactions as two classes of legal and fraudulent transactions. Detecting credit card fraud is based on an analysis of card spending behaviour. There have been some attempts in the past to use different machine learning approaches and feature selection technologies in transaction datasets. The...
classification of credit card transactions is mostly a binary classification problem. Here, the credit card transaction is either a valid transaction (negative class) or a fraudulent transaction (positive class). Detecting fraud is generally considered a data mining classification problem where the goal is to correctly classify credit card transactions as legal or fraudulent.

Many methods have been applied to credit card fraud detection, artificial neural networks, genetic algorithms, support vector machines, often item set mining, decision trees, migrate birds optimization algorithms, and nave base. A comparative analysis of logistic regression and naïveBayes will be conducted. The performance of Bayesian and neural networks can be assessed on credit card fraud data. Decision trees, neural networks, and logistics regression are tested for their applicability in detecting fraud [1].

III. PROPOSED METHOD-

Methodologies are specific to each and every project .Machine learning provides various algorithms for supervised learning basically, supervised learning algorithms are classified into two categories namely classification algorithms and regressionalgorithms. As the dataset only two output labels are classified. The KNN algorithm hinges on this assumption being true enough for the algorithm to be useful. The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other. The KNN algorithm follows the below steps:

1. Load the data.
2. Initialize K to your chosen number of Neighbors.
3. For each example in the data:
   3.1 Calculate the distance between the query example and the current example from the data.
   3.2 Add the distance and the index example to an ordered collection.
4. Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances.
5. Pick the first K entries from the sorted collection.
6. Get the labels of these selected K entries.
7. Return the mode of the K labels.

THE ALGORITHM-
The KNN Algorithm follows the below steps:
1. Load the data.
2. Initialize K to your chosen number of Neighbors.
3. For each example in the data:
   3.1 Calculate the distance between the query example and the current example from the data.
   3.2 Add the distance and the index example to an ordered collection.
4. Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances.
5. Pick the first K entries from the sorted collection.
6. Get the labels of these selected K entries.
7. Return the mode of the K labels.
FEATURE (VARIABLE) SELECTION:

The basis of credit card fraud detection lies in the analysis of cardholder’s spending behaviour. This spending profile is analysed using optimal selection of variables that capture the unique behaviour of a credit card. The profile of both a legitimate and fraudulent transaction tends to be constantly changing. Thus, optimal selection of variables that greatly differentiate both profiles is needed to achieve efficient classification of credit card transactions. The variables that form the card usage profile and techniques used affect the performance of credit card fraud detection systems. These variables are derived from a combination of transaction and past transaction history of a credit card. These variables fall under five main variable types, namely all transaction statistics, regional statistics, merchant type statistics, time-based amount statistics and time-based number of transactions statistics. The variables that fall under all transactions statistics type depict the general card usage profile of the card. The variables under regional statistics type show the spending habits of the card with respect to geographical regions. The variables under merchant statistics show how the usage of the card in different merchant categories. The variables of time-based statistics types identify the usage profile of the card with respect to usage amounts versus time ranges or frequencies of usage versus time ranges.

Most literature focused on cardholder profile rather than card profile. It is evident that a person can operate two or more credit cards for different purposes. Therefore, one can exhibit different spending profile on such cards. In this study, focus is beam on card rather than cardholder because one credit card can only exhibit a unique spending profile while a cardholder can exhibit multiple behaviors on different cards.

Dataset: Initially, the dataset downloaded from Kaggle.com has 31 attributes which comprises of 30 input attributes and 1 output attribute. The 30 input attributes include Time, Amount and 28 variables (v1-v28) and these 28 attributes are unknown as they are confidential and encrypted for security purpose. It contains only numerical (continuous) input variables which are as a result of a Principal Component Analysis (PCA) feature selection transformation resulting to 28 principal components. The details and background information of the features cannot be presented due to confidentiality issues. As it is difficult for the user to enter the unknown 28 variables. The dataset is minimized to lesser columns. The columns that are essential to be in the dataset are selected by finding the correlation of each and every input variable with the output class variable. The values for which the correlation values are higher are collected and considered to form the final dataset. The dataset was modified to form a dataset with 13 input attributes and 1 output attribute. The dataset is highly unbalanced and skewed towards the positive class. The 13 attributes selected for the dataset are Time, Amount, v1, v3, v4, v7, v10, v11, v12, v14, v16, v17, v18 and the output variable is class. All the unknown variables are undergone PCA transformation to hide the confidential information.

Experimental Analysis: Accuracy is considered as one of the measures in predicting the result of an experiment. To check the performance of our model, we use accuracy score. To get the accuracy score, we have to determine the confusion matrix.

Confusion matrix: In the field of machine learning and specifically in the problem of statistical classification, a confusion matrix is also known as an...
error matrix. It is a specific tabular layout that allows visualization of the performance of an algorithm. Typically it is used in supervised learning and in unsupervised learning it is usually called a matching matrix. Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class or vice versa.

![Confusion Matrix](image)

**Figure 4: Confusion Matrix**

**Accuracy:** For a classification problem, the result obtained as either of the class is considered as the correct result by measuring accuracy of the experiment. Accuracy, sensitivity, specificity, precision etc. are the performance evaluators. Accuracy is the ratio of number of correct predictions to the total number of input samples. It works well only if there are equal numbers of samples belonging to each class.

Accuracy = \( \frac{TP}{TP + TN + FP + FN} \)

**ROCCurve:**

A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. ROC analysis provides tools to select possibly optimal models and to discard suboptimal ones independently from (and prior to specifying) the cost context or the class distribution.

![ROC Curve](image)

**Figure 5: ROC Curve**

**Comparative Study:**

**AUC for different algorithms:** Checking the accuracy for different algorithms helps us to obtain the best algorithm in solving our problem. ROC analysis is related in a direct and natural way to cost/benefit analysis of diagnostic decision making.

![AUC for different algorithms](image)

**Figure 6: AUC for different algorithms**

**CONCLUSION:**

Although there are several fraud detection techniques available today but none is able to detect all frauds completely when they are actually happening, they usually detect it after the fraud has been committed. From this model, frauds occurring with credit card can be easily detected. Detecting the frauds in advance by training the machine with previous transaction data enables us to safeguard our details and thereby preventing the loss of money.