International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 5 May 2022, pp: 841-844 www.ijaem.net ISSN: 2395-5252

Credit Card Fraud Identification

Riya Dubey, Siddhant Jain, Parth Sharma, Ms. Bhaivee Singh
[1,2,3,4] B-tech Cse, Galgotias University, Gautam Buddha Nagar, India

---

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 Rossam 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.

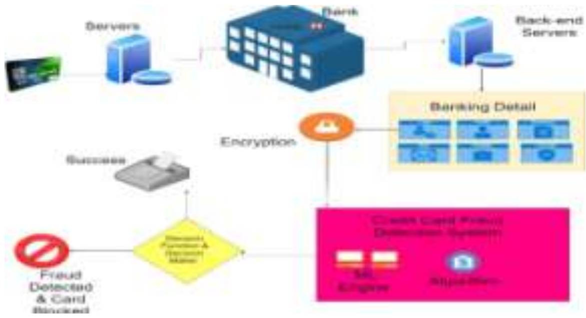


Figure 1: Credit card fraud detection system

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 learningprovidesvariousalgorithmsforsupervisedlearning.Basically,supervisedlearningalgorithms are classified into two categories namely classification algorithms and regressionalgorithms.Asthedatasethasonlytwooutputlabelsitcomesunderclassificationalgorithms. Asthedatasethasonlytwooutputlabelsitcomesunderclassificationalgorithms.

Initially train the dataset with various classification algorithms such as Logistic Regression,K-NearestNeighbours,SimpleVectorMachine,DecisionTreeandRandomForest.Compare accuracy given by each algorithm and choose the algorithm with best accuracy. Forpredicting,Classification algorithm called KNN isused.

TheK-nearestNeighbors(KNN)algorithmisasimple,easy-to-implementsupervisedmachinelearningusedtosolve classificationandregressionproblems .Thek-nearestneighbors(KNN)algorithmisasimple,easy-to-implementsupervisedmachinelearning algorithm that can be used to solve both classification and regression problems .TheKNN algorithm assumes that similar things exist in close proximity. In other words, similarthingsareneartoeachother. TheKNNalgorithmhingesonthisassumptionbeingtrueenoughforthealgorithmtobeuseful.KNNcapturestheideaof similarity(sometimescalleddistance,proximity, or closeness) with some mathematics we might have learned in our childhood—calculating the distance between points on a graph. Distance between two points can becalculated by mathematical formula and there are other ways of calculating distance, and oneway might be preferable depending on the problem we are solving. However, the straight-linedistance (also called the Euclidean distance) is a popular and familiar choice. To select the Kthat's right for your data, we run the KNN algorithm several times with different values of Kand choose the K that reduces the number of errors we encounter while maintaining thealgorithm'sabilitytoaccuratelymakepredictionswhenit'sgivendataithasn'tseenbefore.

Fig.2 Fraud alert

THE ALGORITHM-

TheKNNAlgorithmfollows thebelowsteps:

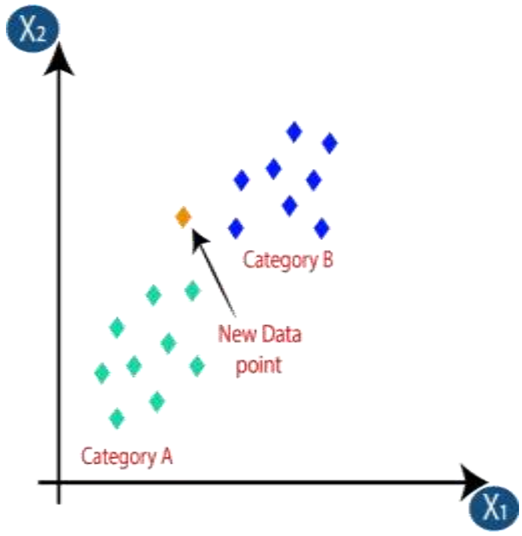
1. Loadthedata.
2. InitializeKtoyourchosennumberofNeighbors.
3. Foreachexampleinthe data

3.1 Calculate the distance between the query example and the current examplefrom thedata.

3.2 Addthedistanceandtheindex exampleto anordered collection.

4. Sort the ordered collection of distances and indices from smallest to largest (inascendingorder)bythedistances.
5. PickthefirstKentries fromthesortedcollection.
6. Getthelabelsof theselectedKentries.
7. Returnthe modeof theKlabel.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 5 May 2022, pp: 841-844 www.ijaem.net ISSN: 2395-5252



FEATURE (VARIABLE) SELECTION:

The basis of credit card fraud detection lies in the analysis ofcardholder's spending behaviour. This spending profile is analysed using optimal selection ofvariables that capture the unique behaviour of a credit card. The profile of both a legitimateandfraudulenttransactiontendstobeconstantly changing.Thus,optimalselectionofvariablesthatgreatly differentiatesbothprofilesisneededtoachieveefficientclassificationof credit card transaction. The variables that form the card usage profile and techniques usedaffect the performance of credit card fraud detection systems. These variables are derivedfromacombinationoftransactionandpasttransactionhistoryofacreditcard.

Thesevariablesfallunderfivemainvariabletypes,namely alltransactionsstatistics,regionalstatistics, merchant type statistics, timebased amount statistics and time-based number oftransactions statistics. The variables that fall under all transactions statistics type depict thegeneral card usage profile of the card. The variables under regional statistics type show thespending habits of the card with taken into account the geographical regions. The variablesundermerchantstatisticstypeshowtheusageofthecardindifferentmerchantcategories.The variables of time based statistics types identify the usage profile of the cards with respecttousageamountsversustimerangesorfrequencies ofusageversustimeranges.

Most literature focused on cardholder profile rather than card profile. It is evident that a person canoperate two or more credit cards for different purposes. Therefore, one can exhibit differentspending profile on such cards. In this study, focus is beamed on card rather than cardholderbecause one credit card can only exhibit a unique spending profile while a cardholder canexhibitmultiplebehaviors on different cards.

Dataset: Initiallythedatasetdownloadedfromthe Kaggle.comhas31attributeswhichcomprises 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 the details of creditcard are usually confidential and encrypted for security purpose. It contains only numerical(continuous) input variables which are as a result of a Principal Component Analysis (PCA)featureselectiontransformationresultingto28principalcomponents.Thedetailsandbackgroundinformationofthefeaturescannotbepresentedduetoconfidentialityissues.Asit is difficult for the user toenter the unknown 28 variables. The dataset is minimized tolesser columns. The columns that are essential to be in the dataset are selected by finding thecorrelation of each and every input variable with the output class variable. The values forwhich 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 attributes.The dataset is highly unbalanced and skewed towards the positive class. The 13 attributesselected for the dataset are Time, Amount,v1,v3,v4,v7,v10,v11,v12,v14,v16,v17,v18 and theoutput variable is class. All the unknown variables are undergone PCA transformation to hidetheconfidential information

Time	Amount	V1	V3	V4	V7	V10	V11	V12	V14	V16	V17	V18	Class
0	149.62	-1.35981	2.536347	1.378155	0.239599	0.090794	-0.5516	-0.6178	-0.31117	-0.4704	0.207971	0.025791	0
0	2.69	1.191857	0.16648	0.448154	-0.0788	-0.16697	1.612727	1.065235	-0.14377	0.463917	-0.1148	-0.18336	0
1	378.66	-1.35835	1.773209	0.37978	0.791461	0.207643	0.624501	0.066084	-0.16595	-2.89008	1.109969	-0.12136	0
1	123.5	-0.96627	1.792993	-0.86329	0.237609	-0.05495	-0.22649	0.178228	-0.28792	-1.05965	-0.68409	1.965775	0
2	69.99	-1.15823	1.548718	0.403034	0.592941	0.753074	-0.82284	0.538196	-1.11967	-0.45145	-0.23703	-0.03819	0
2	3.67	-0.42597	1.141109	-0.16825	0.476201	-0.37141	1.341262	0.359894	-0.13713	0.401726	-0.05813	0.068653	0
4	4.99	1.229658	0.045371	1.202613	-0.00516	-0.09925	-1.41691	-0.15383	0.167372	-0.44359	0.002821	-0.61199	0
7	40.8	-0.64427	1.07438	-0.4922	1.120631	1.249376	-0.61947	0.291474	-1.32387	-0.07613	-1.22213	-0.35822	0
7	93.2	-0.89429	-0.11319	-0.27133	0.370145	-0.41043	-0.70512	-0.11045	0.074355	-0.21008	-0.49977	0.118765	0
9	3.68	-0.33826	1.044367	-0.22219	0.651583	-0.36685	1.017614	0.83639	-0.44352	0.739453	-0.54098	0.476677	0
10	7.8	1.449044	0.91386	-1.37567	-1.42324	1.626659	1.199644	-0.67144	-0.09505	0.001967	0.253415	0.854344	0
10	9.99	0.384978	-0.8743	-0.09402	0.470455	0.309755	-0.25912	-0.32614	0.362832	-0.12945	-0.80998	0.359985	0
10	121.5	1.249999	0.38393	-1.2349	-0.6894	1.323729	0.227666	-0.24268	-0.31763	-0.81561	0.873936	-0.84779	0
11	27.5	1.069374	0.828613	2.71252	-0.09672	0.46023	-0.77366	0.323387	-0.17849	-0.19993	0.124005	-0.9805	0
12	58.8	-2.79185	1.64175	1.767473	-0.42291	1.151087	0.844555	0.792944	-0.73498	-0.30306	-0.15587	0.778265	0
12	15.99	-0.75242	2.057323	-1.46864	-0.60858	0.747731	-0.79398	-0.77041	-1.0666	1.660114	-0.27927	-0.41999	0
12	12.99	1.103215	1.267332	1.289091	-0.58606	-0.26798	-0.45031	0.936708	-0.46865	-0.24663	-0.00921	-0.59591	0
13	0.89	-0.43691	0.924591	-0.72722	0.707642	-0.73798	0.324098	0.277192	-0.2919	1.143174	-0.92871	0.68047	0
14	46.8	-5.40126	1.186305	1.736239	-1.55974	0.345175	0.91723	0.970117	-0.47913	0.472004	-0.72548	0.075081	0
15	5	1.492936	0.454795	-1.43803	-1.08066	1.638076	1.077542	-0.63205	0.052011	-0.16643	0.304241	0.554432	0

Experimental Analysis: Accuracy is considered as one of the measure in predicting the resultof an experiment.A To check the performance of our modelwe use accuracy score. To getthe accuracyscoreinitiallywehaveto determinethe confusion matrix

Confusionmatrix: Inthefield of machine learning and specifically in the problemof statistical classification, a confusion matrix is also known as an

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 5 May 2022, pp: 841-844 www.ijaem.net ISSN: 2395-5252

error matrix. It is a specifictable layout that allows visualization of the performance of an algorithm. Typically it is usedin supervised learning and in unsupervised learning it is usually called a matching matrix.Eachrowofthe matrix representstheinstancesinapredictedclasswhileeachcolumnrepresentstheinstances in an actualclass orvice versa

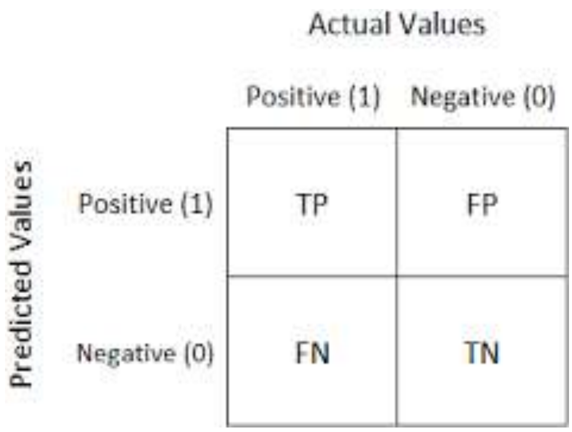


Figure4:ConfusionMatrix

Accuracy: For a classification problem, the result obtained as either of the class is considered asthe correct result by measuring accuracy of the experiment. Accuracy, sensitivity, specificity,precisionetc. are the performance evaluators .Accuracy istheratioofnumber of correctpredictions to the total number of input samples. It works well only if there are equal numberofsamples belongingtoeachclass.

Accuracy= TP.
(TP+TN+FP+FN)

ROCCurve:
A receiveroperatingcharacteristiccurve,or ROCcurve,isagraphicalplotthatillustrates the diagnostic ability of a binary classifier system as its discrimination threshold isvaried.ROCanalysisprovidestoolstoselectpossiblyoptimalmodelsandtodiscard suboptimal ones independently from (and prior to specifying) the cost context or the classdistribution

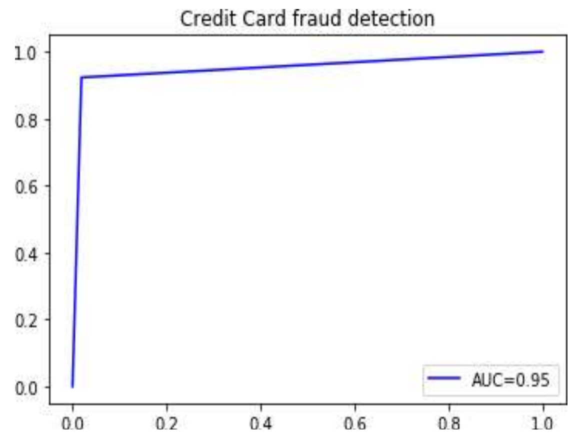


Figure5:ROCCurve

Comparativestudy:
AUC for different algorithms: Checking the accuracy for different algorithms helps us toobtain the best algorithm in solving our problem. ROC analysis is related in a direct andnaturalwayto cost/benefit analysis of diagnosticdecisionmaking

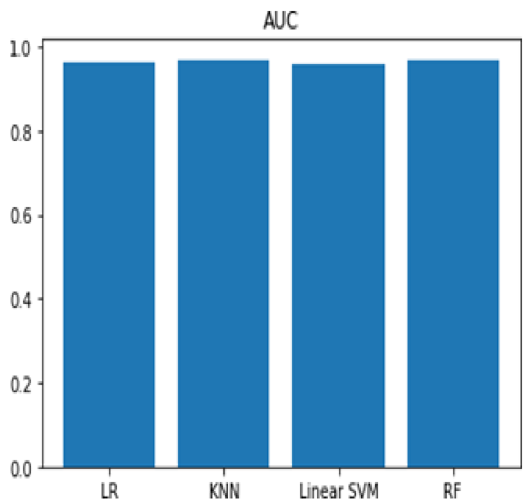


Figure6:AUCfordifferentalgorithms

CONCLUSION:

Although there are several fraud detection techniques available today but none is able todetect all frauds completely when they are actually happening, they usually detect it after thefraud has been committed. From this model, frauds occurring with credit card can be easilydetected. Detecting the frauds in advance by training the machine with previous transactionsdataenables usto safeguard ourdetails andtherebypreventingtheloss ofmoney.