CREDIT CARD FRAUD DETECTION

RICHARD

2K18/SE/106

richard_2k18se106@dtu.ac.in

SHIVOM

2K18/SE/119

shivom_2k18se119@dtu.ac.in

ABSTRACT: It is important that the customers are not charged for the items that have not purchased from their credit cards. These fraudulent transactions are frequently taking place nowadays. These problems can be solved using machine learning techniques.Machine learning is an important topic which can help us solve various problem with ease. Credit card fraud detection is one of them.Here,we will explain the modelling of data set with this fraud detection. We will model past credit card transactions that came out as fraud.This model will help us find that which transaction is fraud or not.Here we will focus on analysing and pre- processing of data set and use detection algorithms such as decision tree,SVM and K-nearest neighbour algorithms on the credit card fraud detection data.

KEYWORDS: credit card fraud detection, SVM, KNN, decision tree, machine learning algorithms

INTRODUCTION: Credit card is small plastic card which is used by customers to purchase items and goods from shops,pay bills etc.Credit cards can be read by ATM and can also be used for internet banking.It is issued by banks to the customers based on the promise that the customer returns the money that he used through the credit card plus the extra charges that the bank has specified within a specified limit of time.Every customer has a unique number on their credit card which the user should keep private and should not tell anyone.Since the last few years,there has been rapid growth in usage of credit cards which also resulted in the rise of fraudulent transactions taking place.A fraud occurs when a person uses a credit card of other individuals without their prior knowledge.This results in depleting of the entire amounts from their credit cards.Here we will try to tackle this problem of credit cards frauds with the help of machine learning. The fraudsters have developed a lot of ways to do frauds with customers.But here,we will use certain algorithms which will help us identify fraud transactions.We will try develop a system which is efficient by the usage of machine learning algorithms.

DATA SETS: Here we have used data sets from kaggle.It contains transaction made by European credit card holders in September 2013. This dataset tells us about transactions that took place in two days where we found out that 492 transactions were fraud out of 284,807 transactions.Our data set is highly unbalanced. Positive class(fraud) accounts for only 0.172% of all transactions. It comprises of only numerical input variables which are result of PCA transformation. Due to confidentiality issues, original features and more background information about the data cannot be provided. V1,V2…..V28 features are principal components that are obtained with PCA. 'Time' and 'amount' are the only features not obtained with PCA. 'Time' feature contains the seconds elapsed between each transaction and the

first transaction in the dataset. 'Amount' is the transaction Amount used for example-dependant cost-senstive learning. 'Class' feature is the response variable and it takes value 1 in case of fraud and 0 otherwise.

CREDIT CARD FRAUD DETECTION METHODS

1.Decision Tree

2.Support vector machines

3.K-nearest neighbour

In our research paper,we will be focussing on these algorithm and explain that how they are used credit card fraud detection.

EXPERIMENTAL DESIGN:

This section shows the details of experiments performed on different models in this study.

Firstly, a performance comparison is made on the same dataset.

Then we explore the relation between legal and fraud transactions in the dataset. Finally, in this section we have showed the whole procedure of execution of this study on a machine using different models and later we discussed there results.

A. Performance measures

Before we describe the experiment, we first introduce the measures we used. Because accuracy rate is not enough to measure the performance of model that we have used. When the data is significantly imbalanced. For instance, a default prediction of all instances into the majority class will also have a high value of accuracy. Therefore, we need to consider other measures. Such as

Precision rate is a measure of the result of prediction and recall rate measures the detection rate of all fraud cases. F-measure is the mean of recall and precision. Auc score which represents the probability that a random positive (green) example is positioned to the right of a random negative (red) example.

Table 1: Confusion Matrix

		Predicted Class	
		Negative	Positive
Actual Class	Negative	True Negative (TN)	False positive (FP)
	Positive	False Negative (FN)	True Positive (TP)

True Positive (TP): These values are correctively predicted positive that means value of both actual class and predicted class are YES.

False Positive (FP): These value of actual class is NO and value f predicted class is YES.

False Negative (FN): These value of actual class is YES and value f predicted class is NO.

$$PRECISION=\frac{TP}{TPxFP}$$

$$RECALL=\frac{TP}{TP+FP}$$

$$F1\text{-}score=2X\frac{precisionXrecall}{Precision+recall}$$

B. Experimental dataset

The dataset we have used has transactions made by European credit cards in September 2013.This dataset contain transactions that happened in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of total transactions.

C. Experiment I

The aim of this experiment is to show how K-Nearest Neighbour algorithm is used to identifying fraud detection. We have used and above mentioned dataset for this experiment. Then 80% transactions of the above dataset are as the training sample, and the rest as the testing sample. The dataset used in this experiment is highly biased toward legit transactions so to nullify it we have used Standard Scaler on test and train samples of the dataset. We have implemented Knn algorithm with value of k as 5 because we see a high error rate for test set when K=1. Hence we can say that model over fits when value of k is 1. For value of K higher than 1, we

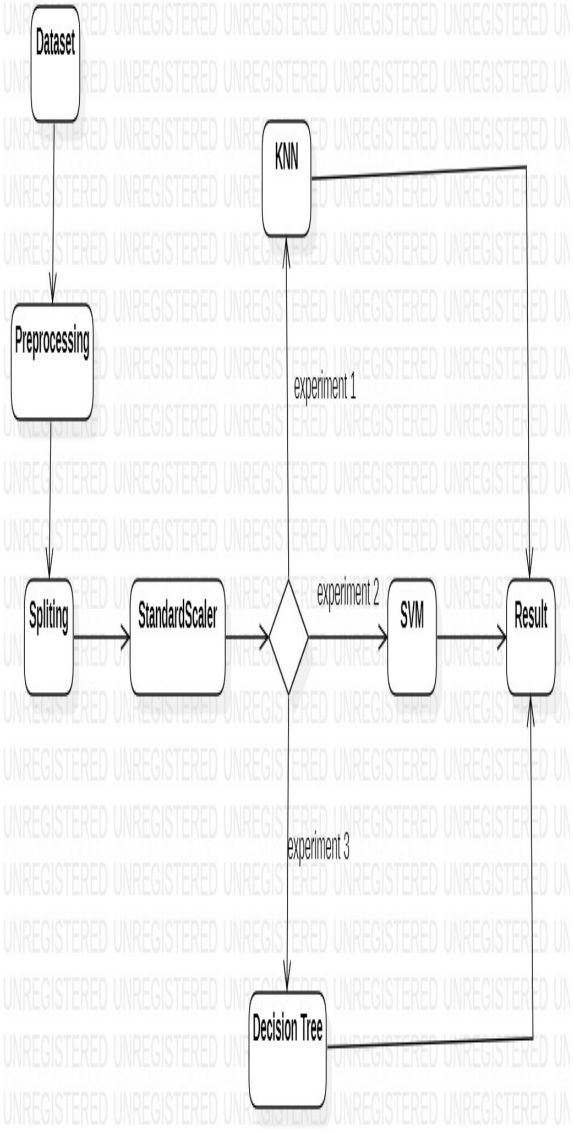
see that the F1 score starts to decrease. We reaches a minimum error rate when value of k is 5.

D. Experiment II

The aim of this experiment is to show how Support vector machine algorithm is used to identifying fraud detection. We have used and above mentioned dataset for this experiment. Then 80% transactions of the above dataset are as the training sample, and the rest as the testing sample. The dataset used in this experiment is highly biased toward legit transactions so to nullify it we have used Standard Scaler on test and train samples of the dataset. We have implemented Support vector machine algorithm (svc) with RBF Kernel Function.

E. Experiment III

The aim of this experiment is to show how decision Trees algorithm is used to identifying fraud detection. . We have used and above mentioned dataset for this experiment. Then 80% transactions of the above dataset are as the training sample, and the rest as the testing sample. The dataset used in this experiment is highly biased toward legit transactions so to nullify it we have used Standard Scaler on test and train samples of the dataset. We have implemented decision Trees algorithm with entropy as the criterion



RESULTS:

We have applied three ml model in form of three different experiments.

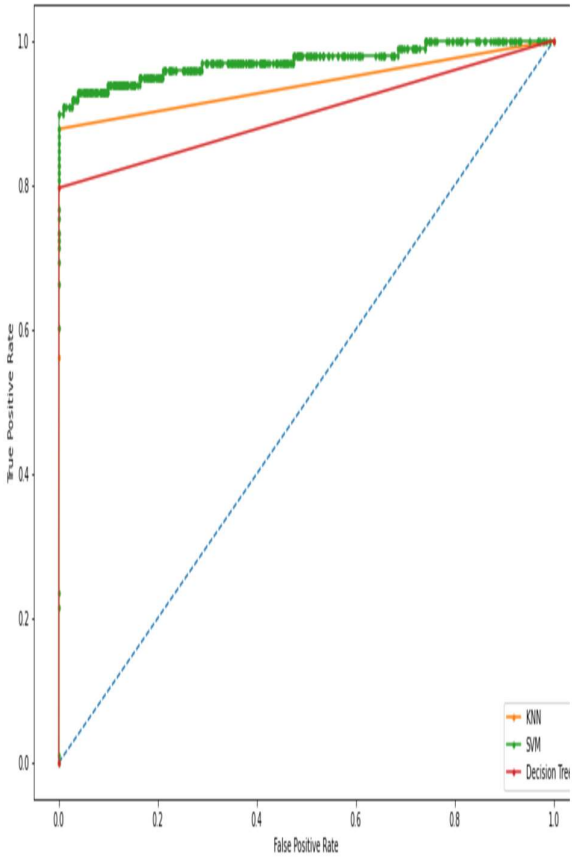
Result from each of the experiment is show in this section.

Table 2 will show the result comparison b/w all the three models.

Models	Accuracy (%)	Recall (%)	precision (%)	F1 score	AUC score
KNN	99.95	79.59	93.975	0.861	0.939
SVM	99.94	69.38	95.774	0.804	0.972
Decision tree	99.92	79.59	78.0	0.787	0.898

Table 2

We have also generated a ROC curve for all the three model:



CONCLUSION: In this paper we developed a method for fraud detection, where customers are grouped based on their transactions and extract behavioural patterns to develop a profile for every cardholder. Then different classifiers are applied on the dataset and rating scores are generated for every type of classifier. After analysing the result mentioned above we came to know that among all the algorithm used in this paper svm came to be the best for anomaly prediction in case of credit card fraud detection with an accuracy of 99.942% and AUC score of 0.972 and it also cover maximum area in roc curve. On the knn and decision tree better recall value than svm. Thus use of svm algorithm for fraud probably in a very short span of time after the transactions has been made. This will eventually prevent the banks and customers from great losses and also will reduce risks.

REFERENCES: [1] What is credit card fraud? definition and meaning. [Online].

[2] Supervised and Unsupervised Machine Learning Algorithms, Machine Learning Mastery, 22-Sep-2016. [Online].

[3] C. Cortes, V. Vapnik **Support vector networks** Machine Learning. 20 (1995), pp. 273-297.

[4] [3] J. Steele and J. Gonzalez, Credit card fraud and ID theft statistics, CreditCards.com. [Online].

[5]IJERT research paper