# A Comparative Study Of Credit Card Fraud Detection Using Machine Learning for United Kingdom Dataset.

Under the Esteemed Guidance of

Dr. M. SUJATHA

(Associate Professor) CSE Department

V. SANDEEP, T. SUPRIYA, P. HIRANMAI

## Abstract:

Financial fraud is an ever growing with far consequences in the financial industry. The growth in E-commerce industry has lead to exponential increase in the UK of credit cards for online purchases and consequently they have been surge in the fraud related to it. In recent years, for banks has become very difficult for detecting the fraud in credit card system. Machine learning plays a vital role for detecting the credit card frauds in the transactions. For predictive these transactions, banks make use of various methodologies, past data has been collected and new features are been used for enhancing the predictive power. The performance of fraud detecting in credit card transactions is greatly affected by the sampling approach on dataset, selection of variables and detection techniques can be used. This paper investigates the performance of fraud Logistic Regression, Random Forest, Decision Tree. The dataset of credit cards can be collected from Kaggle and it contains 2,84,808 credit card transactions of an European bank dataset. It considers that fraud transactions as the positive classes and genuine ones as negative class. The dataset is highly imbalanced. The performance of fraud detection in credit card transactions is greatly affected by the sampling approach on dataset, selection of variables and detection techniques used. A hybrid technique of under sampling and over sampling is carried out on the skewed data. The three techniques are applied on the raw and the preprocessed data. The work is implemented in Python. The performance of the techniques is evaluated based on accuracy, sensitivity, specificity, precision, Matthews's co-relation coefficient and balanced classifications. Logistic Regression is best among all the algorithms.

**Keywords**— Credit card fraud, Credit Card Fraud detection methods ,Electronic Commerce, Hidden Markov Model(HMM) , Logistic Regression., Random Forest.

## 1. Introduction

In day to day life credit cards are used for purchasing goods and services with the help of virtual card for online transaction or physical card for offline transaction. In a physical-card based purchase, the card holder presents his card physically to a merchant for making a payment. To carry out fraudulent transactions in this kind of


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

purchase, an attacker has to steal the credit card. If the card holder does not realize the loss of card, it can lead to a substantial financial loss to the credit card company. In online payment mode, attackers need only little information for doing fraudulent transaction (secure code, card number, expiration date etc.).In this purchase method, mainly transactions will be done through Internet or telephone. To commit fraud in these types of purchases, a fraudster simply needs to know the card details. Most of the time, the genuine card holder is not aware that someone else has seen or stolen his card information. The only way to detect this kind of fraud is to analyze the pending patterns on every card and to figure out any inconsistency with respect to the "usual" spending patterns. Fraud detection based on the analysis of existing purchased data of card holder is a promising way to reduce the rate of successful credit card frauds. Since humans tend to exhibit specific behaviorist profiles, every card holder can be represented by a set of patterns containing information about the typical purchase category, the times since the last purchase, the amount of money spent, etc. Deviation from such patterns is a potential threat to the system.

## 2. RELATEDWORK:

Lakshmi's ,Selvani, Deepthi,Kavila(2018) demonstrate that efficiency classification models to credit card fraud detection problems and the author proposed three models Logistic Regression, Random Forest, Neural Network, among these three neural network and Logistic Regression out performs than the Decision Tree . M.JIslametal (2017) proposed a probability theory frame work for making decision under classificant k-Nearest Neighbour, Navie-Bayes.

After reviewing the Bayesian theory, Naive-Based algorithm and these algorithms mainly used

For detecting the fraud. Y.Sachin and E.Doman (2018) has cited the search for credit card fraud detection and used several classification methods took a major role. In this work they have included decision tree and SVM to decrease the risks of banks. They have suggested ANN (Artificial Neural Network) and Logistic Regression classification models are helpful to improve the performance in detecting the cards. Here the training datasets become more biased and the efficiency of all models decreased in catching fraudulent transactions.

# 3.Methdology

In the methodology step we describe some kind of machine learning algorithms that can used to predict the credit card fraud usage. The following System architecture shows these algorithms and they can describe the class labels (Yes/No), give some accuracy measures.

## 3.1 Problem Definition:

An Ecommerce payment system facilitates the acceptance of electronic payment for online transactions. Ecommerce payments system have been increasingly popular due to the wide spread use of the internet-based shopping and banking. Since in traditional system the fraud is detected only when the billing for the credit card is done It is difficult to


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

prevent or detect the fraudulent transactions.

## 3.2 Proposed Technique:

The proposed technique is used in this frame work , for detecting frauds in credit card system. This proposed may come across with the different classification algorithm such as Random Forest, Linear Regression, Logistic Regression, Decision Tree, KNN (K-Nearest Neighbour). Which algorithm gives suits best and can be adopted by credit card merchants for identifying fraud transactions. The figure shows the architectural techniques for representing the overall system frame work.

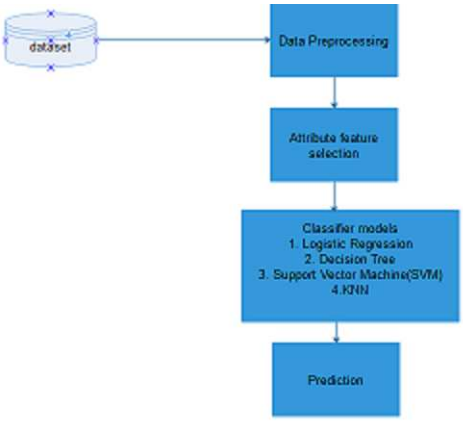


**Fig: System Architecture.**

## 3.3CREDIT CARD FRAUD DETECTION METHODS OR ALGORITHMS:

In order to detect the frauds usage of credit card we are applying some algorithms that predict the class label as “YES/NO”. Different strategies can be used and every strategy having its own methodology and giving some accuracy rate by which we can easily understand every algorithm in proper manner.

Some of the following algorithms are:

I. K- NEAREST NEIGHBOR ALGORITHM

The concept of credit card fraud detection by using a data stream outlier detection algorithm which is based on reverse k-nearest neighbors (SODRNN). The distinct quality of SODRNN algorithm is it needs only one pass of scan. Whereas traditional methods need to scan the database many times, it is not suitable for data stream environment [16].

The performance of KNN algorithm is influenced by three main factors:

•The distance metric used to locate the nearest neighbors.

• The distance rule used to derive a classification from k nearest neighbor.

• The number of neighbors used to classify the new sample.

II.DECISIONTREE

Decision Trees are statistical data mining technique that express independent attributes and a dependent attributes logically AND in a


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

tree shaped structure. Classification rules, extracted from decision trees, are IF-THEN expressions and all the tests have to succeed if each rule is to be generated [8].The work demonstrates the advantages of applying the data mining techniques including decision Trees and SVMs to the credit card fraud detection problem for the purpose of reducing the bank's risk. The results show that the proposed system decision tree it shows the accuracy rates in predicting the class labels'.

## III. Logistic Regression

Logistic Regression is one of the classification algorithm, used to predict a binary values in a given set of independent variables (1 / 0, Yes / No, True / False). To represent binary / categorical values, dummy variables are used. For the purpose of special case in the logistic regression is a linear Regression, when the resulting into the different values.

## IV. SUPPORT VECTOR MACHINE (SVM)

The basic idea of SVM classification algorithm is to construct a hyper plane as the decision plane which making the distance between the positive and negative mode maximum [18]. The strength of SVMs comes from two important properties they possess - kernel representation and margin optimization. Kernels, such as radial basis function (RBF) kernel, can be used to learn complex regions. A kernel function represents the dot product of projections of two data points in a high dimensional feature space. In SVMs, the classification function is a hyper-plane separating the different classes of data. The basic technique finds the smallest hyper sphere in the kernel space that contains all training instances, and then determines on which side of hyper sphere a test instance lies. If a test instance lies outside the hyper sphere, it is confirmed to be suspicion SVM can have better prediction performance than BPN (Back propagation network) in predicting the future data. But in large data BPN has a good performance.

## 4. Experiment Analysis and Results:

In the section we demonstrate that accuracy rates, results of each and every algorithm we mentioned as above. Every algorithm will give some accuracy rates and we can understand the algorithm easily. The accuracy rate shoes the performance of the algorithms and this will detect the frauds easily.

The following figure shows the comparison of all classification algorithms:

**This figure shows that comparison of al l classification algorithms .The X- axis denotes that all the algorithms and the Y - axis denotes that all the accuracy rates.**


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

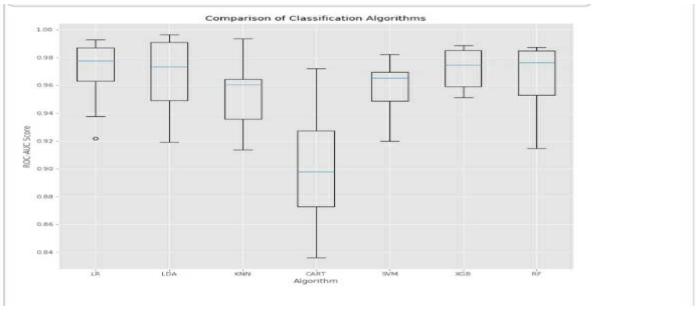


Fig: comparison of All classification Algorithms based on Accuracy

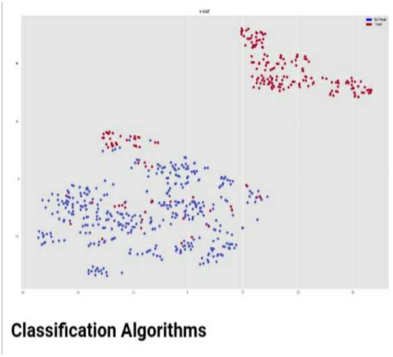


Fig: clustering of All Algorithms

The performance of the NavieBayes algorithm can be described as:

Cross Validation mean Score of NavieBayes :83.6%

Model Accuracy: 84.7%

Confusion Matrix:

[ [ 170270 20205]

[38005 152466]]

**Table: Different Performance metrics of NavieBayes classifier:**

| Precision | Recall | f1 – score | Support |
|---|---|---|---|
| 0.85 | 01904771 | 0.82 | 0.89 |
| O. 84 | 11904771 | .0.88 | 0.80 |
| avg/total | 0.85 | 0.85 | 0.85 |

The performance of the Logistic Regression is described as:

CrossValidationmeanScoreofNavieBayes99.6%

ModelAccuracy99.7%

Confusion Matrix:

[[17527020278]

[38505152568]]

**Table: Different Performance metrics of Logistic Regression :**

| Precision | Recall | f1-score | Support |
|---|---|---|---|
| 0.95 | 01904992 | 0.96 | 0.98 |
| 0.90 | 11904998 | 0.96 | 1.0 |
| Avg /total | 0.95 | 0.96 | 0.99 |

The performance of the NavieBayes algorithm can be described as:

Cross Validation mean Score of K-Nearest Neighbor (KNN):83.6%

Model Accuracy: 98.7%

Confusion Matrix:

[[74687020705]

[6840558666]]

**Table : Different Performance metrics of KNN Algorithm:**

| Precision | Recall | f1-Score | Support |
|---|---|---|---|
| 0.87 | 01904994 | 0.96 | 0.98 |



https://sites.google.com/site/ijcsis/
ISSN 1947-5500

| | | | |
|---|---|---|---|
| 0.88 | 1190499 8 | 0.97 | 0.95 |
| avg/total | 0.98 | 0.97 | 0.96 |

The performance of the NavieBayes algorithm can be described as:

Cross Validation mean Score of K-Nearest Neighbor (KNN):83.6%

Model Accuracy: 98.7%

Confusion Matrix:

[[74687020705]

[6840558666]]

**Table: Different performance metrics of KNN Algorithm:**

| Precisio n | Recall | f1-Score | Suppor t |
|---|---|---|---|
| 0.87 | 0190499 4 | 0.96 | 0.98 |
| 0.88 | 1190499 8 | 0.97 | 0.95 |
| avg/total | 0.98 | 0.97 | 0.96 |

**Table: Different performance metrics of all classification algorithms:**

| Algorithm | Accuracy |
|---|---|
| LR | 0.969945(0.022522) |
| KNN | 0.966647(0.027772) |
| Naive Bayes | 0.952231(0.023558) |
| Decision Tree | 0.952231(0.023558) |
| SVM | 0.966171(0.022672 |
| RF | 0.96654(0.025672 |

The following figure shows the decision tree for detecting the frauds.

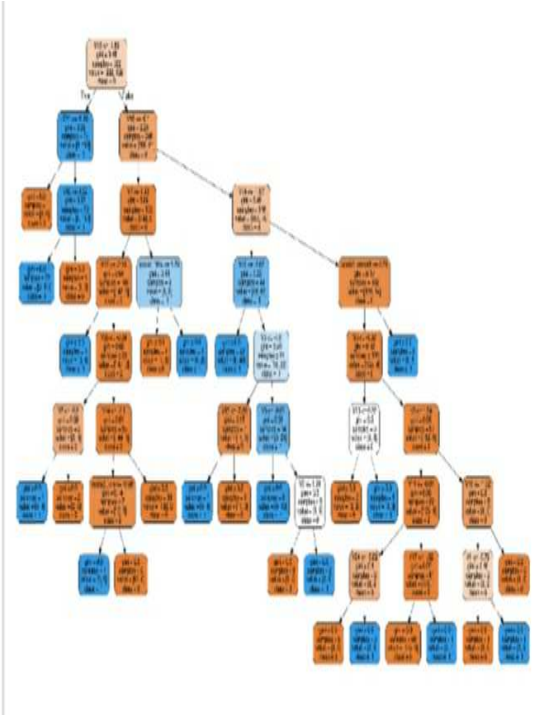

Fig: Decision tree for detecting the frauds.

## 4. Conclusion

Currently, Credit card risk monitoring system is one of the key tasks for the merchant banks, organization to improve merchants risk management level in an automatic, scientific and adequate way. There are many ways of detection of credit card fraud. If one of these or combination of algorithm is applied into bank credit card fraud detection system, the probability of fraud transactions can be predicted soon after credit card transactions by the banks and a series of anti- fraud strategies can be adopted to prevent banks from great losses before and reduce risks. This paper gives contribution towards the effective ways of credit card fraudulent detection. In this study, three


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

classification methods were used to a deep analysis of the credit cards history business information and have built the fraud detecting models. We present our work and demonstrate the advantages of the logistic regression and decision tree to the credit card fraud detection, for the purpose of reducing the bank's risk. The results show that the proposed classifier neural networks and logistic regression approaches outperform decision tree in solving the problem under investigation.

REFERENCES:

[1]Sandeep Pratap Singh, Shiv Shankar P.Shukla,Nitin Rakesh and Vipin Tyagi "Problem Reduction In Online Payment System Using Support vector machine" International Journal of Managing Information Technology (IJMIT) Vol.3, No.3, August 2011

[3] Sam Maes, Karl Tuyls, Bram Vanschoenwinkel, Bernard Manderick,"Credit card fraud detection using Bayesian and neural networks, "Interactive image-guided neurosurgery, pp.261-270,2018.

[4 ] Sam Maes, Karl Tuyls, and Bram, "Credit Card Fraud Detection using Bayesian &Neural Network", 2018.

[5 ] Raghavendra Patidar, Lokesh Sharma, "Credit Card Fraud Detection Using Neural Network", International Journal of Soft Computing and Engineering (IJSCE)ISSN: 2231-2307, Volume-1, Issue,NCAI2011, June 2011.

[6]Krishna Kumar Tripathi1, Mahesh A. Pavaska "Survey on Credit Card Fraud Detection Methods "International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, Volume 2, Issue 11, November 2016).

[7 ] Venkata Ganji, Siva Naga Prasad Mannem, "Credit card fraud detection using anti-k nearest neighbor algorithm", International Journal on Computer Science and Engineering (IJCSE), Vol. 4 No. 06 June 2015.

[8 ] Y. Sahin and E. Duman, "Detecting Credit Card Fraud by Decision Trees and Support Vector Machines", International Multi conference of Engineers and computer scientists March, 2015.

[9] Joseph King-Fung Pun, "Improving Credit Card Fraud Detection using a Meta-Learning Strategy", Chemical Engineering and Applied Chemistry University of Toronto 2018.

[10] lakshims, selavni depthi ETAL" Credit card fraud detection using Machine learning ".

[11] Aihua Shen1, Rencheng Tong1, Yaoche Deng21 "Application of Classification Models on Credit Card Fraud Detection" University of the Chinese Academy of Sciences, chennai.


https://sites.google.com/site/ijcsis/
ISSN 1947-5500