International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 5, Issue 10, pp 45-50, October-2021

https://doi.org/10.47001/IRJIET/2021.510009

Comparative Study on Different Fraud Detection Techniques

[1]Kavya Murali, [2]Prof. Indu K B, [3]Prof. Pragisha K

[1,2,3]Department of Computer Science and Engineering, LBS College of Engineering, Kasaragod, Kerala, India

***Abstract* - Frauds are known to be dynamic and have no patterns, hence they are not easy to identify. Fraudsters use recent technological advancements to their advantage. They have somehow bypass security checks, leading to the loss of millions of dollars. Analyzing and detecting unusual activities using data mining techniques is one way of tracing fraudulent transactions. The work presented in this paper provides an empirical study and analysis of supervised learning techniques, that Logistic regression, K nearest neighbours, SVM, Random forest, Naïve Bayes , on a bench mark credit card transaction dataset. The performance results have been evaluated and compared to identify the best predictive technique. The techniques have been used to detect whether a given transaction is fraudulent or not.**

I. INTRODUCTION

We all know in today's world online scams are increasing day by day. They are known by different names like e-crime, cyber crime etc. Whatever names they are known they create disguise problems to people which they affect. They create financial problems and many victims are also discovered. They are different kinds of online scams like credit card, lottery scams, phishing, career oppurtunities, romance fraud etc. Online fraud detection is mainly based on machine learning algorithms. Machine learning technique gives the computer to learn without being explicitly programmed. ML provides efficient result to extract knowledge by constructing models from dataset. There are different types of machine learning algorithms they are supervised machine learning algorithm, unsupervised machine learning algorithm and reinforcement machine learning algorithm. Despite the variety of techniques fraudsters use to cover their tracks very few manage to completely obfuscate the fraud traces they leave behind. Unsupervised techniques aim to detect anomalies in unlabelled datasets by grouping events according to their features. Clustering is a commonly used technique and more recently, auto encoders have also proven useful for feature extraction. Auto encoders learn a lower dimensional latent representation of the input data and then learn to reconstruct it. Supervised techniques require the dataset to contain a label indicating whether an event is fraudulent or not. However, labelled datasets are scarce with only a few high quality public datasets available. The main aim of this work is to automatically detect frauds in the credit card transactions. The detection of fraud is a very complex computational task. The properties of a good fraud detection system are

1. It should identify the fraud accurately.
2. It should not classify any genuine transaction as fraud.
3. It should find the frauds quickly.

Data's for the credit card fraud detection was taken from kaggle. Different fraud detection methods are given below.

In section 2.1 it shows how SVM is used in credit card fraud detection. Fig 2.0 also gives us a detailed idea about how to construct a hyper plane for linearly separable data, and how its applied in SVM. This paper also gives detailed study of the pros and cons of SVM. In section 2.2 it give a detailed description of another important technique Naïve Bayes and how it is applied in credit card fraud detection. This section also defines the importance of each attribute by using a well known data set of a golf play in Fig 2.1. It also describes the pros and cons of Naïve Bayes. In section 2.3 we discuss about the most important machine learning technique Random Forest. It also shows the structure of the Random Forest using a Fig 2.3. The Fig 2.3 shows the detail picture of Random Forest and how it is used in a dataset and how to predict fraud and non fraud. It also shows a confusion matrix of fraud and non fraud in Fig 2.4. It also figures out the main pros of Random Forest. In section 2.4 it describes about Logistic regression and how it is used in fraud detection. It also shows a graph showing the fraud and non fraud cases. In section 2.5 it describes another detection technique KNN and how it is used. Section 3 compares all the detection techniques and in section 4 we just try to find which detection technique is the best.

II. LITERATURE SURVEY

The recent papers evaluate the performance of certain data mining tools for predicting the frauds.

2.1 SVM (Support vector machine)

SVM is mainly used for classification or regression. It allows sophiscated non linear issues like credit card fraud

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 5, Issue 10, pp 45-50, October-2021

https://doi.org/10.47001/IRJIET/2021.510009

detection to be solved using linear classification without increasing computational complexity. SVM[1][3] finds a hyper plane that creates boundary between types of data. In 2-dimensional space, this hyper-plane is nothing but a line. In SVM, we plot each data item in the dataset in an N-dimensional space, where N is the number of attributes in the data. SVM can only perform binary classification. To perform SVM on multi-class problems, we can create a binary classifier for each class of the data. The two results of each classifier will be

a) The data point belongs to that class OR
b) The data point does not belong to that class.

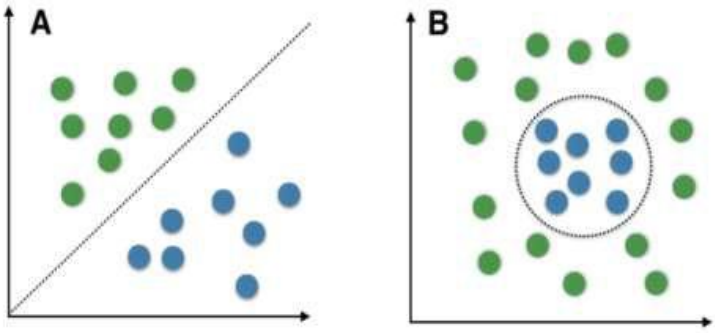

Fig 2.0

SVM works very well for linearly separable data. Linearly Separable Data is any data that can be plotted in a graph and can be separated into classes using a straight line. Kernelized SVM is used for non-linearly separable data. A kernel is nothing a measure of similarity between data points. There are various kernel functions available,

a) Radial Basis Function Kernel (RBF)
b) Polynomial Kernel

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. The strength of SVM comes from two important properties they possess kernel representation and marginal optimization. RBF are used to learn about complex regions. A kernel function represents the dot product of projections of two data points in a high dimensional feature space. The basic technique finds the smallest hyper sphere in the kernel space that contains all the training instances, and then determines on which side of hyper sphere a test instance lies. If a test instances lies outside the hyper sphere it is confirmed to be suspicion. SVM finds a special kind of linear model, the maximum margin hyper plane and it classifies all the training instances correctly by separating them into correct classes through a hyper plan. The maximum margin hyper plane is the one that gives the greatest separation between the classes. The instances that are nearest to the maximum margin hyper plane are called support vectors. In credit card fraud detection, for each test instance, it determines if the test instance falls within the learned region. Then if the test instance falls within the learned region, it is declared as normal, else it is declared as anomalous.

Pros of Kernelized SVM:

- They perform very well on a range of datasets.
- They work well for both high and low dimensional data.

Cons of Kernelized SVM:

- Efficiency (running time and memory usage) decreases as size of training set increases.
- Does not provide direct probability estimator.
- Difficult to interpret why a prediction was made.

2.2. NAIVE BAYES

Naive Bayes classifiers [4] are a collection of classification algorithms based on Bayes Theorem. Consider a fictional dataset in Fig 2.1 that describes the weather conditions for playing a game of golf. Given the weather conditions, each tuple classifies the conditions as fit (“Yes”) or unfit (“No”) for playing golf.

	Predictors				Response
	Outlook	Temperature	Humidity	Wind	Class Play=Yes Play=No
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

Fig 2.1

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 5, Issue 10, pp 45-50, October-2021

https://doi.org/10.47001/IRJIET/2021.510009

The dataset is divided into two parts, namely

a) Feature matrix
b) Response vector

Feature matrix contains all the vectors (rows) of dataset in which each vector consists of the value of dependent features. In above dataset, features are "Outlook", "Temperature", "Humidity" and "Windy". Response vector contains the value of class variable (prediction or output) for each row of feature matrix. In above dataset, the class variable name is "Play golf". The fundamental Naive Bayes assumption is that each feature makes an:

a) Independent
b) Equal contribution to the outcome

We assume that no pair of features is dependent. For example, the temperature being "Hot" has nothing to do with the humidity or the outlook being "Rainy" has no effect on the winds. Hence, the features are assumed to be independent. Secondly, each feature is given the same weight (or importance). For example, knowing only temperature and humidity alone can't predict the outcome accurately. None of the attributes is irrelevant and assumed to be contributing equally to the outcome.

Pros of Naïve Bayes

- When assumption of independent predictors holds true, a Naive Bayes classifier performs better as compared to other models.
- Naive Bayes requires a small amount of training data to estimate the test data. So, the training period is less.
- Naive Bayes is also easy to implement.

Cons of Naïve Bayes

- Main limitation of Naive Bayes is the assumption of independent predictors. Naive Bayes implicitly assumes that all the attributes are mutually independent. In real life, it is almost impossible that we get a set of predictors which are completely independent.
- If categorical variable has a category in test data set, which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as Zero Frequency.

Naïve Byes is also used in credit card fraud detection. Fig 2.2 shows the performance scores under different K features. The accuracy is an average score to show how much percentage we get the right prediction. Precision is how much probability we get a sample with true if it's tested positive. Higher precision means with transaction identified as fraud by this model model we have higher correct rate, higher recall means, if the transaction is fraud, then we have higher chance to identify correctly.

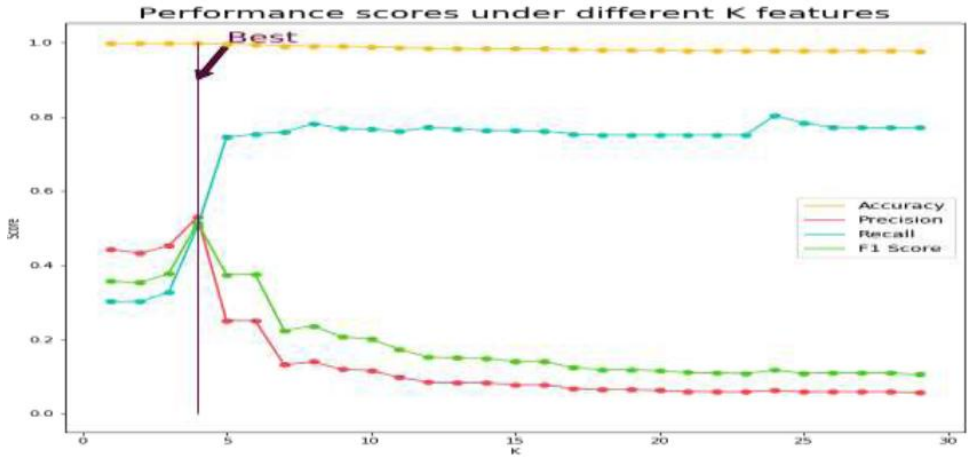

Fig 2.2

2.3 Random Forest

Random forest is a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. Random Forest [1][2] selects the best feature rather than the most important feature among a random subset of data resulting in a better model. Thus having a binary classification of fraud i.e. positive case (value 1) and non fraud i.e negative case (value 0) for the target category in the transaction amount. Random forest is

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 5, Issue 10, pp 45-50, October-2021

https://doi.org/10.47001/IRJIET/2021.510009

based on ensemble learning. Ensemble learning is an algorithm where the predictions are derived by assembling or bagging different models or similar model multiple times. The random forest algorithm works in a similar way and uses multiple algorithm i.e. multiple decision trees, resulting in a forest of trees, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification.

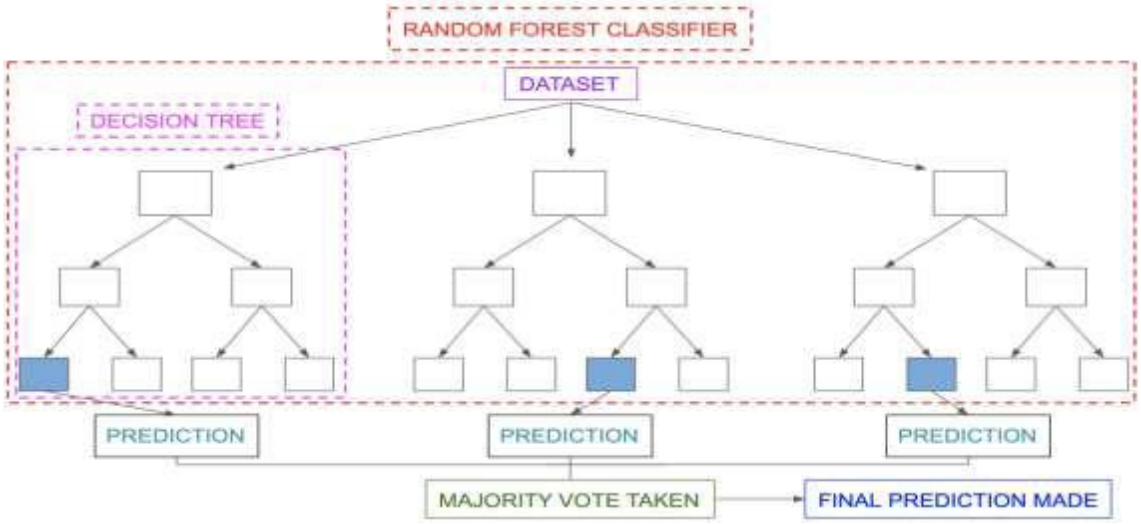

Fig 2.3

Pros of Random Forest

- The random forest algorithm is not biased and depends on multiple trees where each tree is trained separately based on the data, therefore biased ness is reduced overall.
- It's a very stable algorithm. Even if a new data point is introduced in the dataset it doesn't affect the overall algorithm rather affect the only a single tree.
- It works well when one has both categorical and numerical features.
- The random forest algorithm also works well when data possess missing values, or when it's not been scaled properly.

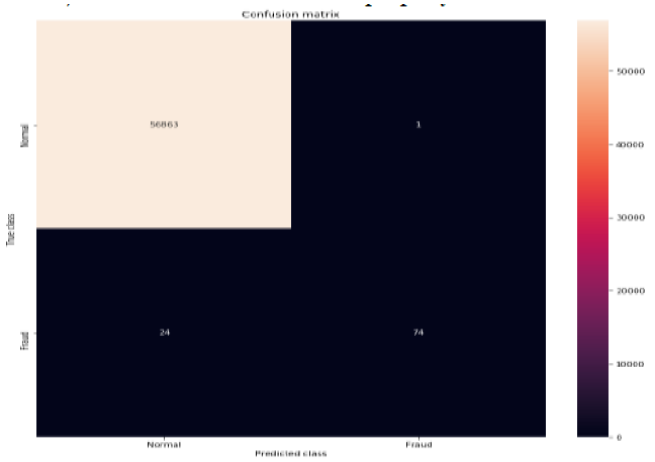

Fig 2.4

Fig 2.4 confusion matrix shows the fraud in the credit card transactions.

2.4 Logistic Regression

Logistic regression [3] is a classification algorithm. It is used to predict a binary outcome based on a set of independent variables. A binary outcome is one where there are only two possible scenarios either the event happens (1) or it does not happen (0). Independent variables are those variables or factors which may influence the outcome (or dependent:

a) Continuous
b) Discrete, ordinal
c) Discrete, nominal

Different types of logistic regression

a) Binary logistic regression
b) Multinomial logistic regression
c) Ordinal logistic regression

Binary logistic regression: It statistical technique used to predict the relationship between the dependent variable (Y) and the independent variable (X), where the dependent variable is binary in nature. Multinomial logistic regression: It is used when you have one categorical dependent variable with two or more unordered levels (i.e two or more discrete outcomes). It is very similar to logistic regression except that here you can have more than two possible outcomes.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 5, Issue 10, pp 45-50, October-2021

https://doi.org/10.47001/IRJIET/2021.510009

Ordinal logistic regression: It is used when the dependent variable (Y) is ordered (i.e., ordinal). The dependent variable has a meaningful order and more than two categories or levels.

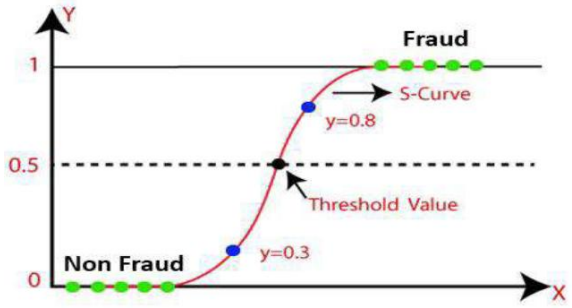

Fig 2.5

From Fig 2.5 we can analyze that the fraud class takes the value "1", while the non fraud class takes the value "0". A threshold of 0.5 is used to differentiate between the two classes, as shown in Figure.

Pros of Logistic Regression

- Logistic regression is easier to implement than linear regression and is very efficient to train.
- It makes no assumptions about the distributions of classes in the feature space.
- It can easily be extended to multiple classes (multinomial regression).
- It is very efficient for classifying unknown records.

Cons of Logistic Regression

- Logistic regression fails to predict a continuous outcome.
- Logistic regression assumes linearity between the predicted (dependent) variable and the predictor (independent) variables.
- Logistic regression may not be accurate if the sample size is too small.

2.5 K Nearest Neighbour (KNN)

K-NN is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN [2] algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

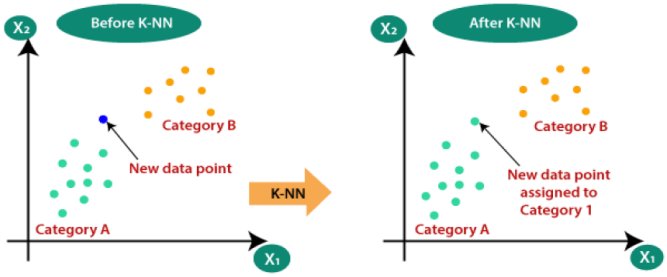

Fig 2.6

Fraud detection techniques based on K-NN technique in credit card require various distance measures is defined in between two data instances. While implementing KNN technique, we classify any of the input transaction by estimating the nearest point to the new input transactions. Fig 2.6 shows the pictorial view of how the new data points are assigned. As the neighbours of particular objects give votes, so with the maximum votes of that objects neighbour, the objects gets classified ie the object is classified to the k the nearest neighbor which is most common to it. This rule just continues and holds the complete training set during the learning stage of the K nearest neighbor algorithm and assigns each and every query through which a particular class is represented by the maximum number of its k-nearest neighbours in the training data set. The KNN rule is the most rudimentary form of KNN when the k"s value is equal to 1. If the nearest neighbour is identified to be a fraudulent transaction, then it is termed as a fraud one KNN, we classify any incoming transaction by calculating nearest point to new incoming transaction. If the nearest neighbour is fraudulent, then the transaction is classified as fraudulent and if the nearest neighbour is legal, then it is classified as legal, the KNN achieves consistently high performance, without apriori assumptions about the distributions from which the training examples are drawn.

III. COMPARISON

Fig 2.7 compares the different machine learning techniques to detect the fraud in credit card fraud detection.

What other Data Scientists got

Method Used	Frauds	Genuines	MCC
Naïve Bayes	83.130	97.730	0.219
Decision Tree	81.098	99.951	0.775
Random Forest	42.683	99.988	0.604
Gradient Boosted Tree	81.098	99.936	0.746
Decision Stump	66.870	99.963	0.711
Random Tree	32.520	99.982	0.497
Deep Learning	81.504	99.956	0.787
Neural Network	82.317	99.966	0.812
Multi Layer Perceptron	80.894	99.966	0.806
Linear Regression	54.065	99.985	0.683
Logistic Regression	79.065	99.962	0.786
Support Vector Machine	79.878	99.972	0.813

Fig 2.7

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 5, Issue 10, pp 45-50, October-2021

https://doi.org/10.47001/IRJIET/2021.510009

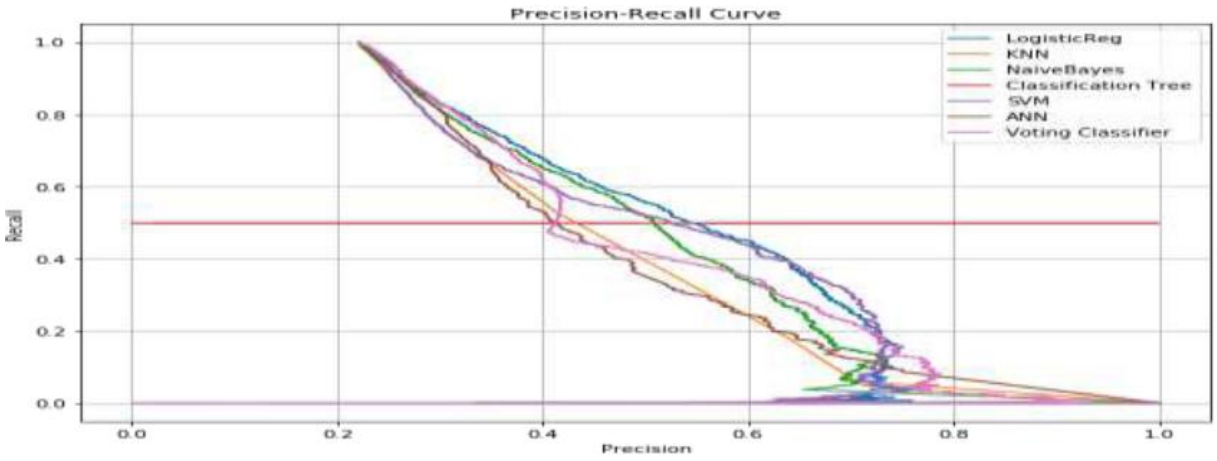

Fig 2.8

PRC comparison of various algorithms is shown in Fig 2.8 By comparing algorithms, a Voting classifier has good accuracy but when we draw PRC, it shows that Logistic regression has good Precision-Recall value at threshold 0.5 so, while changing threshold values, it improves the Precision and Recall values.

IV. CONCLUSION

We have discussed about different machine learning techniques and compared the process and its features. Each technique has its own advantages and disadvantages. When compared to different machine learning techniques random forest is having high accuracy and recall. F1 score of random forest s highest. SVM provides high accuracy and is expensive. In SVM, if a test instance lies outside the hyper sphere, it is confirmed to be suspicion transaction. K-nearest neighbour imposes high processing speed and is expensive. Naïve Bayes classification is done by applying Bayes rule to calculate the probability of the correct class shows good performance.

REFERENCES

[1] Enhanced credit card fraud detection based on SVM-recursive feature elimination and hyper-parameters optimization Naoufal Rtayli*, Nourddine Enneya Laboratory of Informatics Systems and Optimization, Ibn Tofail University, Kenitra, Morocco.

[2] Fraudulent Transaction Detection in Credit Card by Applying Ensemble Machine Learning. Debachudamani Prusti Department of Computer Science and Engineering National Institute of Technology Rourkela Rourkela, India

[3] Data mining for credit card fraud: A comparative study Siddhartha Bhattacharyya Sanjeev Jha , Kurian Tharakunnel , J. Christopher Westland Data mining for credit card fraud: A comparative study Siddhartha Bhattacharyya Sanjeev Jha Kurian Tharakunnel , J. Christopher Westland.

[4] Spam Detection in Social Media Employing Machine Learning Tool for Text Mining.Zakia Zaman Department of Computer Science and Engineering Bangladesh University of Engineering and Technology. Sadia Sharmin Department of Computer Science and Engineering Bangladesh University of Engineering and Technology.

Citation of this Article:

Kavya Murali, Prof. Indu K B, Prof. Pragisha K, "Comparative Study on Different Fraud Detection Techniques" Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 5, Issue 10, pp 45-50, October 2021. Article DOI https://doi.org/10.47001/IRJIET/2021.510009

****

