SURVEY ON CREDIT CARD SECURITY SYSTEM FOR BANK TRANSACTION USING NAÏVE BAYESIAN AND RANDOM FOREST

Ram Sundar G[1], Joe Franklin J[2], Karthikeyan K M[3], Arun N[4]

[1]*Associate Professor, Dept. of Computer Science and Engineering, KPRIET, Tamilnadu, India*
[2,3,4]*Student, Dept. of Computer Science and Engineering, KPRIET, Tamilnadu, India*

Abstract - *The use of credit cards is prevalent in modern day environment. The importance of the machine learning and data science cannot be explained in detail. In this paper I have tried to illustrate the modelling of data set using machine learning classification with credit card fraud detection as the base. The credit card fraud detection problem includes modeling past credit card transactions with the knowledge of ones that turn out to be fraud. The model is then used to identity whether a new transaction is fraud or not. But it is obvious that the number of credit card fraud cases is increasing constantly in spite of the chip cards integration worldwide with existing protection systems. Because of this the problem of fraud detection is very important now. In the paper the general description of the developed fraud detection system and the comparisons between models based on artificial intelligence are given. In the last section the results of evaluative testing and corresponding conclusions are considered.*

Key Words: **Naive Bayesian, Random forest, credit card safety, credit card fraud, Fraud detection**

1. INTRODUCTION

Usage of credit cards in modern day society is prevalent. But financial fraud is also occurring in spite of the chip cards worldwide integration and existing protection systems in other related fields. In processing systems most software developers are trying to improve existing fraud detection methods. These methods are majorly rules based models. Such models allow rules describing transactions to bank employees to determine that are suspicious. But new types of fraud appear quickly and the number of transactions per day is large. Therefore, to create corresponding rules in time it is very difficult to track new types of fraud. Increase in number of employees is required significantly. Artificial intelligence is used to avoid such problems. But this task is very special and are not acceptable because of complex models and authorization time limits. The use of Random forest is suitable for this type of detection, but results from previous research showed that some input data (attributes of transaction) representation method should be used for effective classification. The Naïve bayseian model was developed for transaction monitoring by bank employees. This model allows provision of fast analysis of transactions by attributes. In this paper a general description of the developed credit card fraud detection system, the Naive Bayesian Classifier and Random forest with the data representation method are considered. The credit card fraud problem damages inflicted are growing rapidly year after year. In the year 2014 alone, it was estimated that the total global monetary loss was in 16.31 billion dollars – accumulation of damages from card issuers, acquiring banks and merchants. Although new technology is introduced to the public and being mandated by the government agency to replace the old magnetic stripe to EMV (Europay-MasterCard-VISA).Moreover, the perpetrator has no relationship with the cardholder or issuer, and has no intention of informing the card owner of the lost card and making repayments for the transactions made.

2. MOTIVATION:

Credit card fraud transpires when a perpetrator uses somebody's credit card for personal gain and sometimes in absolute secrecy or anonymity; even the issuing banks are unconscious that the card is being utilized. Moreover, the perpetrator has no relationship with the cardholder or issuer, and has no intention of informing the card owner of the lost card and making repayments for the transactions made. The intention of this study is to fully explore the effectiveness of utilizing the credit card transaction logs to differentiate anomalous from legitimate transactions. Moreover the proposed methods are only primarily focused on behaviour of card holder's transaction and sends only fraud message to the processing server to stop the transaction except few servers others doesn't intimate the card holder about the fraudlent and he is unaware of it, So we decided to develop an system to detect the fault transaction happening currently because if a user is doing transaction of 10k for 10 days and on 11th day if he does transaction of 20k the system will consider it as fraud thus send the message to processing server as fraud and to stop the transaction. It will also send a message to user mobile number regarding fraud happened. We implement that message with OTP so that the user if he only does the transaction can enter the OTP on processing server and can continue the process.

3. BACKGROUND:

In this section, we introduce Algorithm technique and their common model.

3.1 Algorithm:

Algorithm techniques in this project introduced are Naïve Bayesian and Random forest model, they process the given data that is collected from various sources and transforms into learning model by processing the collected data sets. The first stage to solve a problem or program is only by algorithm. Algorithm process the data set and it keeps on learning from the given dataset and responds to the outlier data. It does work only on the conditions are satisfied and stops when there is a outlier in condition. It is said to been expansion from pseudo code. Algorithms can be implemented by programs, Algorithms are said to be the predecessor to do a program. Classification algorithms predict one or more discrete variables, based on the other attributes in the dataset. Regression algorithms predict one or more continuous numeric variables, such as profit or loss, based on other attributes in the dataset**.** Segmentation algorithms divide data into groups, or clusters, of items that have similar properties. Association algorithms find correlations between different attributes in a dataset. The most common application of this kind of algorithm is for creating association rules, which can be used in a market basket analysis. Sequence analysis algorithms summarize frequent sequences or episodes in data, such as a series of clicks in a web site, or a series of log events preceding machine maintenance.

3.2 Naive Bayesian:

The supervised learning method Bayesian Classification represents as well as a statistical method for classification. Determining probabilities of the outcomes allows us to captures uncertainty about the model in a principled way by underlying probabilistic model and can solve predictive and diagnostic problems. The powerful and straight forward algorithm for the classification task is the Naïve Bayesian Classifier. Even if we are working with some attributes with millions of records on a data set, it is suggested to try Naive Bayes approach. When we use it for textual data analysis Naive Bayes classifier gives great results such as Natural Language Processing. The Bayes Theorem is inherited by Naive Bayes classifier. For each class it predicts membership probabilities such as the probability that given data point or record belongs to a particular class. The most likely class is the class with the highest probability considered.

P(Hypo|Multievi) = P(E1| Hypo)* P(E2|Hypo) ……*P(En|Hypo) * P(Hypo) / P(Multievi)

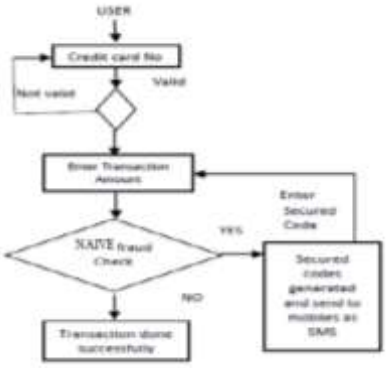

Fig -1: Naïve Bayesian

3.3 RANDOM FOREST:

We use random forest as a classifier in our experiment. The decision tree models in data mining is popular to their simplification in algorithm and flexibility handling different data attribute types. However, specific training data is possibly sensitive on single-tree model. Ensemble methods can solve these problems by combining a group of individual decisions in some way and are more accurate than single classifiers. It one of ensemble methods, is a combination of multiple tree predictors such that each tree depends on a random independent dataset and all trees in the forest are of same distribution. The capacity of random forest not only depends on the strength of individual tree but also correlation between different trees. The stronger the strength of single tree and less the correlation between different tress, the better the performance of random forest. The variation of trees comes from their randomness which involves bootstrapped samples and randomly selects a subset of data attributes. Random forest is still robust to noise and outliers although there are possibly exist some mislabeled instances in our dataset,. We introduce two kinds of random forests, named as random forest 1 and random forest 2, which are different in their base classifiers (i.e., a tree in random forest).For readability, some notations are introduced here. Considering a given dataset D with m examples (i.e. $|D| = m$), we denote: $D = \{(\boldsymbol{xa}, ya)\}$, $a = 1, ..., m$, where $\boldsymbol{xa} \in X$ is an instance in the n-dimensional feature space $X = \{f1, f2, ..., fn\}$ and $ya \in Y = \{0, 1\}$ is the class label associated with instance $\boldsymbol{xa}$.

accuracy=TruePos+ TrueNeg/ TruePos+ FalsePos+ TrueNeg+ FalseNeg
precision=TruePos / TruePos+ FalsePos
recall=TruePos/TruePos+ FalseNeg
FalseNeg– measure =2 × precision × recall/ precision + recall
Intervention =FalsePos /TrueNeg+ FalsePos

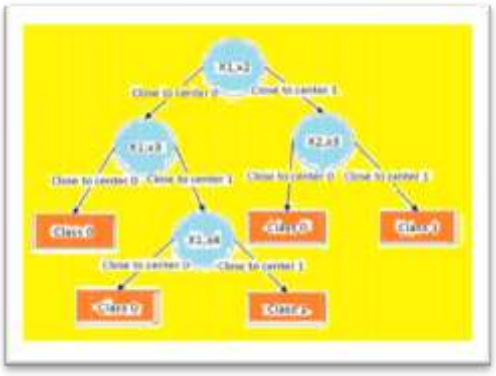

Fig -1: Random Forest

4. LITERATURE REVIEW:

Numerous literatures pertaining to anomaly or noise detection have been published already and are available for public usage. Regardless of the domain, it has been proven that these techniques are effective in obtaining the goal – to identify and differentiate anomaly from the normal instances in the dataset.

In the paper of Chandola they presented an overview of the algorithms that can handle anomaly detection, utilization of these algorithms is still on a case to case basis. Not only have they reviewed previous related works from other literatures, they have also given some insights on the possible hindrances that the researcher might experience in handling such topic. Moreover, the proponents cited other possible applications of these detection algorithms with their advantages and disadvantages. Overall, the paper is a good general background reference in understanding the available anomaly detection techniques.

A similar research domain was presented by Haiying Maand Xin Li where they used an advanced data-mining algorithm which will dynamically produce the ideal attribute combination – the Genetic Algorithm. This algorithm has the ability to perform natural selection of the dataset's attributes heuristically with respect to the chosen fitness-function. By applying this algorithm, it was able to generate a 78%identification rate.

Another similar domain literature is from Wen-Fang Yuand Na Wang where they examined the Outlier data-mining approach to identify fraudulent data-points from the dataset. In this concept, fraud is manifested as an isolated point in the vector space and it could appear independently or some what included in a small group of clustered data-points. Their technique produces 89.4% accuracy when the outlier threshold is set to 12.

Lastly, Duman had successfully implemented a credit card fraud detection system in a Turkish bank using the Migrating Birds Optimization (MBO) algorithm. The paper used the saved limit ratio (SLR) as their primary performance criterion and was fortified by the True Positive Rate (TPR) value. Their MBO algorithm obtained 93.9% for SLR and91.45% for TPR; the selection of this algorithm was further verified via t-tests. The paper also provided some insights on the current fraud detection techniques, option in handling instance distribution of fraudulent and normal transactions, and observations in using common algorithms in this field.

According to Gartner (2014), analytics is categorized into four types: descriptive, diagnostic, predictive, and prescriptive. Descriptive analytics describe what happens in a situation. Analytics later evolved into diagnostic analytics, which tried to address the cause of events. Today, predictive analytics attempt to provide answers to what could possibly happen in the future. The future of analytics is prescriptive analytics, an area still under development. A prescriptive analytics system is a system that can suggest various decision options related to its observations.

Raj & Portia (2011) comparative survey found that each machine learning algorithm adopted for credit card fraud prevention and detection has its various strengths and weaknesses. The strengths of NN identified by Sahinet. al. (2011), Carneiro, Mishra, Ryman-Tubband Raj & Portia (2011) include a high degree of accuracy in detection of hidden patterns, learning, process speed, flexibility and performance. Moreover, the weaknesses of NN include higher false positive rates, and issues related to interpretability of data, model parameter settings, and improper selection issues in a large data set, as well as overall longer

processing times and memory requirements. It is, therefore, imperative that an effective PAT vendor solution adopts the most effective machine learning algorithm to detect and prevent credit card fraud. However, an effective PAT vendor solution for credit card fraud prevention cannot be selected based on machine learning algorithms alone.

Collard (2011) suggested the need for the fraud detection industry to develop and implement enhanced capabilities solutions that provide predictive analytics, pattern recognition, complex event processing, content analytics, name matching, and business rules. A solution must be effective, coherent, agile, flexible, and multi-layered in order to counter fraud effectively. However, existing packaged solutions do not have many of these capabilities. Thus, Collard advocates for the improvement of fraud detection capabilities in neural network-based solutions through the use of more advanced functionalities, such as business rules management systems (BRMS). BRMS can enhance better decisions making, reduce the false positive rate, improve customer experiences, and lower total cost of ownership. Collard also noted that organizations may be resistant to upgrading to more effective alternative fraud detection packages due to an increased total cost of ownership, a preference for black box/closed systems solutions and the ensuing need to change current operational workflow dictated by more advanced solutions.

5. THE VARIOUS INFORMATION RETRIEVAL SCHEMES THAT HAS BEEN DONE BY OTHER AUTHORS ARE STUDIED BELOW:

	Author	Concept	Advantage
Analysis on credit card Fraud detection	S. Benson Edwin Raj, A. Annie Portia	Comparing and analysing some of good techniques that have been used in detecting credit card fraud like Fusion of Dempster Shafer, Bayesian learning, Hidden markov model and artificial neural networks.	Accuracy on Fuzzy, Darwin, Dempster and Bayesian theories in terms of true positive and false positive and also founded that speed is very fast to process these techniques than Hidden Markov Model which shows low accuracy.
Credit card fraud detection based on Transaction Behaviour	John Richard D.Kho, Larry A.Vea	The intention of this project is to fully explore the effectiveness in utilizing the credit card transaction logs to differentiate anomalous from legitimate transactions that helps in predicting the correct classification of input data set.	This method suggest building a model based on the spending behaviour of the card holders and using it to detect anomalous transactions. This system reduces phone and SMS cost shouldered by banks.
The use of predictive Analysis technology to detect credit card fraud	Kosemani Temitayo Hafiz, Dr. Shaun Aghili, Dr.PavolZavarsky	This research paper focuses on side by side comparision of five major credit card predictive analytic vendor solutions adopted. The study depicts the important features and capabilities that should be present in an effective and efficient fraud loss prevention technologies.	The inability of PAT vendor solutions to be 100% effective model and algorithm issues, transactions, data sources.
Retrieval Techniques	Taylor, Francis	Concepts are inferred by the transitive closure of the concept matrix based on fuzzy logic.	This system has implemented a fuzzy information retrieval system called Expert based on the proposed method using Turbo C version 2.0 on a PC/AT
Credit card fraud detection with neural networks	Sushmito Ghosh, Douglas L.Reilly	Using data from a credit card issuer, a neural network based fraud detection system was trained on a large sample of labelled credit card account transactions and tested on a holdout data set that consisted of all account activity over a subsequent two-month period of time.	The feasibility study demonstrated that due to its ability to detect fraudulent patterns on credit card accounts, it is possible to achieve a reduction of from 20% to 40% in total fraud losses, at significantly reduced caseload for human review.

Analysis of Credit Card Fraud Detection Methods	V.Dheepa1 ,Dr.R.Dhanapa	Three approaches to fraud detection are presented. The clustering model, the probability density estimation method and the model based on Bayesian networks. This paper investigates the usefulness of applying different approaches to a problem of Credit card fraud detection.	One aim of this study is to identify the user model that best identifies fraud cases. The models are compared in terms of their performances. To improve the fraud detection system, the combination of the three presented methods could be beneficial.
Credit card fraud detection using neural network and geo location	Aman Gulati, Prakash Dubey, MdFuzailC, Jasmine Norman and Mangayarkarasi R	The proposed system presents a methodology which facilitates the detection of fraudulent exchanges while they are being processed, this is achieved by means of Behaviour and Locational Analysis(Neural Logic) which considers a cardholder's way of managing money and spending pattern. A deviation from such a pattern will then lead to the system classifying it as suspicious transaction and will then be handled accordingly.	In this paper, we present a credit card fraud detection system which works on neural networks seem to detect up to 80% accuracy with sample transaction data(real data results may vary by a little difference)
Credit Card Fraud Detection via Kernel-Based Supervised Hashing	Zhenchuan Li, Guanjun Liu, Shuo Wang, Shiyang Xuan, and Changjun Jiang	The performance of KSH on a real user-related dataset including more than three million transactions from a financial company in China. Experimental results show that KSH achieves the similar performance with the state-of-the art method. More importantly, KSH can offer explainable information that can help investigators decide if a transaction is fraud or not, while these state-of-the-art methods cannot offer these information. Though KNN can also offer these information, KSH is better than it on performance	This paper introduces kernel-based supervised hash method for the first time to deal with fraud detection problem which can provide more explicable information about the suspicious transactions. These information can assist professional investigators take the best measures for fraud alerts. The conclusion that KSH perform well in fraud detection problem, and ANN method is feasible in this research field
Credit Card Fraud Detection Using Machine Learning As Data Mining Technique	Ong Shu Yee, Saravanan Sagadevan and Nurul Hashimah Ahamed Hassain Malim	The combination of machine learning and data mining techniques were able to identify the genuine and non-genuine transactions by learning the patterns of the data. This paper discusses the supervised based classification using Bayesian network classifiers namely K2, Tree Augmented Naïve Bayes (TAN), and Naïve Bayes, logistics and J48 classifiers.	After preprocessing the dataset using normalization and Principal Component Analysis, all the classifiers achieved more than 95.0% accuracy compared to results attained before preprocessing the dataset.

6. REFERENCES:

1) S. Benson Edwin Raj, A. Annie Portia," Analysis on Credit Card Fraud Detection Methods" International Conference on Computer, Communication and Electrical Technology – ICCCET2011, 18th & 19th March, 2011.

2) John Richard D. Kho, Larry A. Vea "Credit Card Fraud Detection Based on Transaction Behavior" Proc. of the

3) 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 2017

4) Kosemani Temitayo Hafiz, Dr. Shaun Aghili, Dr.Pavol Zavarsky "The Use of Predictive Analytics Technology to Detect Credit Card Fraud" "Credit card fraud and interaction debit card statistics - canadian issued cards.," 2014.

5) Sushmito Ghosh and Douglas L. Reilly "Credit Card Fraud Detection with a Neural-Network" SPIE-the International Society of Optical Engineers, 726, 552-558.

6) V.Dheepa, Dr.R.Dhanapal "Analysis of Credit Card Fraud Detection Methods" International Journal of Recent Trends in Engineering, Vol 2, No. 3, November 2009.

7) Aman Gulati, Prakash Dubey, MdFuzail C, Jasmine Norman and Mangayar karasi R "Credit card fraud detection using neural network and Geo location" IOP Conf. Series: Materials Science and Engineering 263 (2017) 042039.

8) Zhen chuan Li, Guanjun Liu, Shuo Wang, Shiyang Xuan, and Chang jun Jiang "Credit Card Fraud Detection via Kernel-Based Supervised Hashing" 2018 IEEE Smart World, Ubiquitous Intelligence& Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovations.

9) Ong Shu Yee, Saravanan Sagadevan and Nurul Hashimah Ahamed Hassain Malim "Credit Card Fraud Detection Using Machine Learning As Data Mining Technique" Journal of Telecommunication, Electronic and Computer Engineering e-ISSN: 2289-8131 Vol. 10 No. 1-4.

10) Shiyang Xuan, Guanjun Liu, Zhenchuan Li, Lutao Zheng, Shuo Wang, Changjun Jiang "Random Forest for Credit Card Fraud Detection" 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNS)