ENHANCED DETECTION OF ANAMOLY ACTIVITIES IN CREDIT CARD SYSTEMS USING SUPERVISED MACHINE LEARNING TECHNIQUES

Deepa.A[1], Elakkiya.S[2], Porselvi.C[3], Senthil.P[4]

[1,2,3]Student, [4]Assistant Professor
Department of Information Technology
Gojan school of Business and Technology

***Abstract*: Web based business is the most useful answer for grow the client base and accomplish the biggest stage with a tiny speculation. The quick development in the E-Commerce has drastically expanded Mastercards use for online buys and it initiated explode in the Visa misrepresentation. For both online just as normal buy Visa turned into the most well-known method of installment, misrepresentation cases associated with it are likewise emerging. The false exchanges are mistaken for authentic exchanges and the basic example coordinating with strategies are not frequently enough to distinguish those fakes precisely. Proficient misrepresentation location framework execution got basic for all Visa giving banks to limit their misfortunes. Current methods dependent on Artificial Intelligence, Data mining, Fuzzy rationale, Machine learning, Sequence Alignment, Genetic Programming and so forth, are advanced in recognizing different Mastercard deceitful exchanges. These methodologies surely lead to a productive Visa extortion recognition framework. This task presents a study of different strategies utilized in Mastercard misrepresentation discovery instruments and assesses every philosophy dependent on certain plan standards.**

Index Terms: Credit Card, Anomaly Detection, Machine Learning, Hidden Markov Model.

I. INTRODUCTION

As we know that a credit card is a small and handy plastic card that is issued by the bank that contains the unique identification such as signature that authorized the person to purchase goods and services on credit and the charges for which will be billed periodically. The information that is stored on the credit card can be read by the Automatic Teller Machines (ATM's), store readers, banks and also used in online Internet Banking System. One of the most important aspects about the credit card is that, it contains a unique card number and the security of the credit card depends on the privacy of the credit card number which is confidential. Due to the rapid growth of credit card transactions has led to a considerable increase in fraudulent activities.

Credit card fraud is an extensive term for the theft and fraud committed using credit card as a fraudulent source of funds in the given transactions. In general, the statistical methods and the data mining algorithms can be used to solve this fraud detection problem. The large number of credit card fraud detection system are based on artificial intelligence, meta learning and pattern matching. The main aim of the genetic algorithm is to obtain the better solution so as to remove the fraud.

The main goal is to develop efficient and secure electronic payment system to detect whether a transaction is fraudulent or not. Here in this paper, we are going to talk about credit card fraud and the measures to detect the fraud. Credit card fraud arises when one person uses other persons' card for their personal use without the knowledge of the card holder. When a card is captured, or stolen or lost, it is used by the fraudsters until the entire available limit of the credit card is depleted. Therefore, we need a way out, which minimizes the total available limit on the credit card which is more prominent to fraud. It aims in minimizing the false alerts using genetic algorithms where a set of interval valued parameters are optimized. Thus the Genetic Algorithm will cause a better solution to such problems. The Importance is given on developing efficient and secure electronic payment system for detecting the fraudulent transactions.

II.TYPES OF CREDIT CARD FRAUD

Fraudulent activities are attempted in many ways such as credit card frauds, frauds through phone calls, computer Intrusions, stealing credit cards, counterfeit fraud, NRI issues etc.

1) Credit Card Fraud: Credit card fraud, in general is of two types: Offline and On-line fraud. Offline fraud can be committed in many ways such as using a stolen physical card, scanning a card like original card or duplicating electronic information of the existing card at any place. Since swiping a card at merchant's place does not require at any a manual signature PIN or a card imprint, On-line frauds are possible even remotely which can take place via internet or telecommunication.

2) Frauds through phone calls: Telephone calls can be used to commit other forms of fraud such as extracting details of card from card holder in the guise of bank authorities. In general, credit card holders, merchants and communication service provider are the

primary preys. That's why the bank regularly warn consumers periodically not to disclose their passwords, cvv number etc., even to bank authorities.

3) Computer Intrusion: The act of breachin privacy without authorization or authentication is defined as Intrusion. It attempting to access and manipulate information purposely in an unauthorized manner.

4) Counterfeit card Fraud: It is a white collar crime that focuses on creating a fake card which holds the details of fully functional original card. This fraud is committed through skimming.

5) CNP Fraud: It is termed as Credit Not Present fraud. This kind of crime can be committed if criminal knows expiry date and the intended account number for card, without being possession of physical card. Any purchase of goods and services using credit cards can be possibly done in two modes: online and offline transaction mode. Every purchase does not require a physical card. Sometimes virtual transactions are also possible. Using physical card based purchase, after purchase, card holder has to swipe his card at the merchant counter for the amount he purchased. The swiping of card requires a hardware known as EMC (Euro pay, MasterCard and Visa) machine. If tin case, the card is stolen, it is crucial to report to police or card issuing company about the loss of card to quash fraudulent activities. The situation becomes worst, if the card holder does not realize loss of card. In the second method of purchasing i.e. online, transactions generally happen through internet by checking the person credentials with the details furnished about a credit card (such as16 digit credit card number, expiry date, CVV number and name of card holder. In order to build a fraud detection mechanism, Data mining and analytics will be the best solution and there are two popular big data technologies that put into practice:

Supervised and unsupervised learning. Data mining and analytics is used to uncover hidden pattern and behaviour, correlations among variables of the dataset.

III.EXISTING SYSTEM

Most of the time, the genuine cardholder is not aware that someone else has seen or stolen his card information. The only way to detect this kind of fraud is to analyse the spending 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 purchase data of cardholder is a promising way to reduce the rate of successful credit card frauds. Since humans tend to exhibit specific behaviourist profiles, every cardholder can be represented by a set of patterns containing information about the typical purchase category, the time since the last purchase, the amount of money spent, etc.
Deviation from such patterns is a potential threat to the system.

IV.DISADVANTAGES

The main disadvantage of the existing system is the detection occurs only after gets a written complaint. In the existing system there is physical inconvenience exists. The period occurs to detect the fraud will cause so many losses to the card holder. There is no particular security system in the existing so a hacker can easily access others card.

V.PROPOSED SYSTEM

Credit card fake recognition which is finished utilizing ML. This strategy is utilized to identify different suspicious exercises on layaway card [2]. It keeps up a database, where past records of exchanges are spared and any strange exchange whenever completed [5], which varies a lot from the past records.

The points of interest of bought things in single exchanges are commonly obscure to any Credit card Fraud Detection System running [7]. Either at the bank that issues charge cards to the cardholders or at the vendor site where products will be bought.
The usage procedures of ML so as to distinguish misrepresentation exchange through credit cards dataset [8]t. It make groups of preparing set and recognize the fraud.

VI.ADVANTAGES

The execution of every grouping calculation was expanded while 10 traits are utilized. From the above trial results, it can likewise be said that utilizing all the 14 qualities for forecast are not all that helpful. It diminishes the exhibition of the classifier as it contains unessential traits. Utilizing the property determination technique for evacuating the superfluous characteristics, it expands the classifiers' exhibition, and calculated relapse demonstrated better execution in forecast .We can find the most accurate detection using this technique. This is redious the tedious work of an employee in the bank.

VII.DATA PREPROCESSING

A. WHY PRE-PROCESSING?

Big data analytics contribute to many fields such as public sector, Education, banking, insurance services and fraud detection. Insurance companies, credit card and phone companies have applied data analytics to perform fraud detection and prevention for decades. Hence there is a need for framework to excavate big data for policing fraudulent activities. The credit card transactions used here are accessed from authenticated online source. Every user may possess more than one card and each card usage is considered as unique profile since user may be using each card for particular purpose.

Therefore, in this paper, we use the credit card transaction dataset which contains a dependent variable that classifies either the customer transaction is fraudulent or not. Further based on the dataset, a classifier model is developed for each for the algorithms (Support Vector Machine) on the training dataset and the remaining test data is tested. The accuracy for each algorithm is calculated from the results obtained. Finally the results obtained are compared. Practical data, in hand are generally Incomplete which may lack the presence of attribute values of interest, or containing only high level data, not in detail. Noisy data or outliers mean the exceptional data which do not fit into the general characteristics of the most data. Inconsistent data means datasets that contain naming inconsistencies, data format or data discrepancies.

B. CHALLENGES IN DATA PRE-PROCESSING

There are many methods available to perform the following steps in data pre-processing.

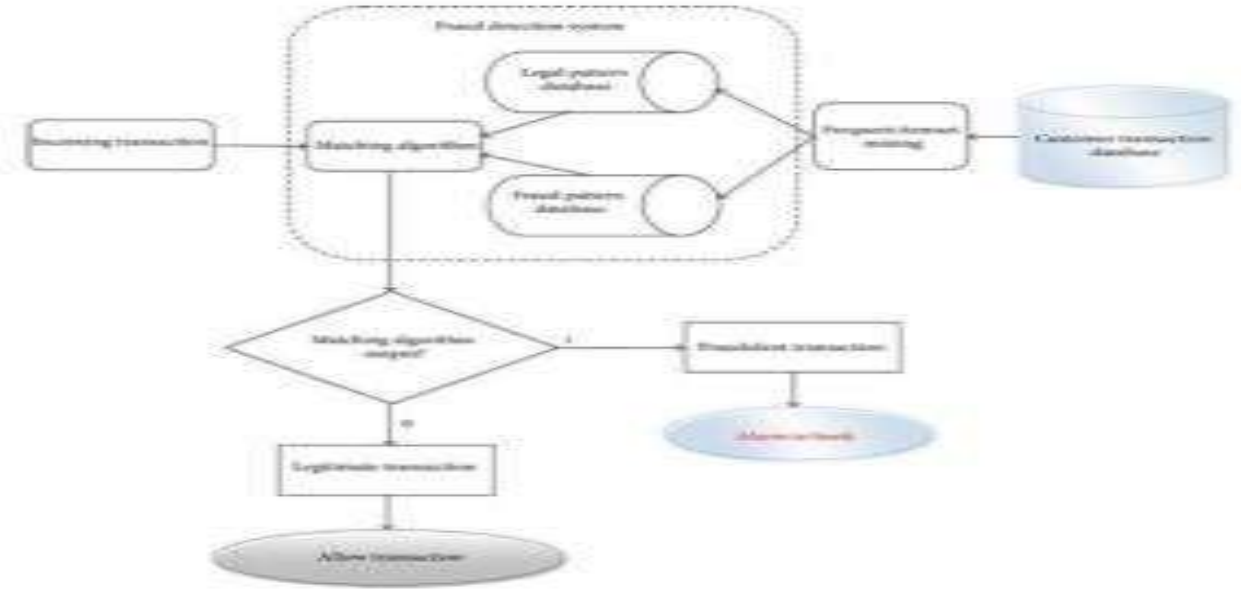
- Data cleaning: Involves filling lost/blank values, even out noisy data that shows variance from the actuals, identifying and removing extreme data,
- Data integration: Data from various databases are to be integrated Data transformation: Transforming the data within a specified range in order to normalize the data.
- Data reduction: Reduced presentation of the original data set but exhibiting the characteristics of the actual volume and also producing similar analytical results as actual data volumes do
- Data discretization: part of data reduction which divides wide range of values into discrete intervals and also replaces numerical attributes with nominal ones using concept hierarchy.

The dataset for analysis is obtained from Kaggle datasets. The dataset contains transactions that occurred in two days, where 492 frauds out of 284,807 transactions are identified. The dataset is highly skewed and imbalanced data, whereas the positive class (frauds) account for 0.172% of all transactions.

The data set contains a total of 30 numerical input variables out of which, 28 variables are the result of a PCA transformation. Due to high confidentiality issues, the original features about the data are not exposed in the website. Features such as V1, V2 ... V28 are the transformed and principal components obtained using PCA transformation technique except the variables 'Time' and 'Amount'. 'Time' is defined as the interval elapsed between each transaction and the first transaction of the dataset. 'Amount' is the spent amount for some purchase. Apart from 30 variable, there is a categorical attribute "class" takes value 1 in case of fraudulent transaction and 0 otherwise.

C. DATA FLOW DIAGRAM

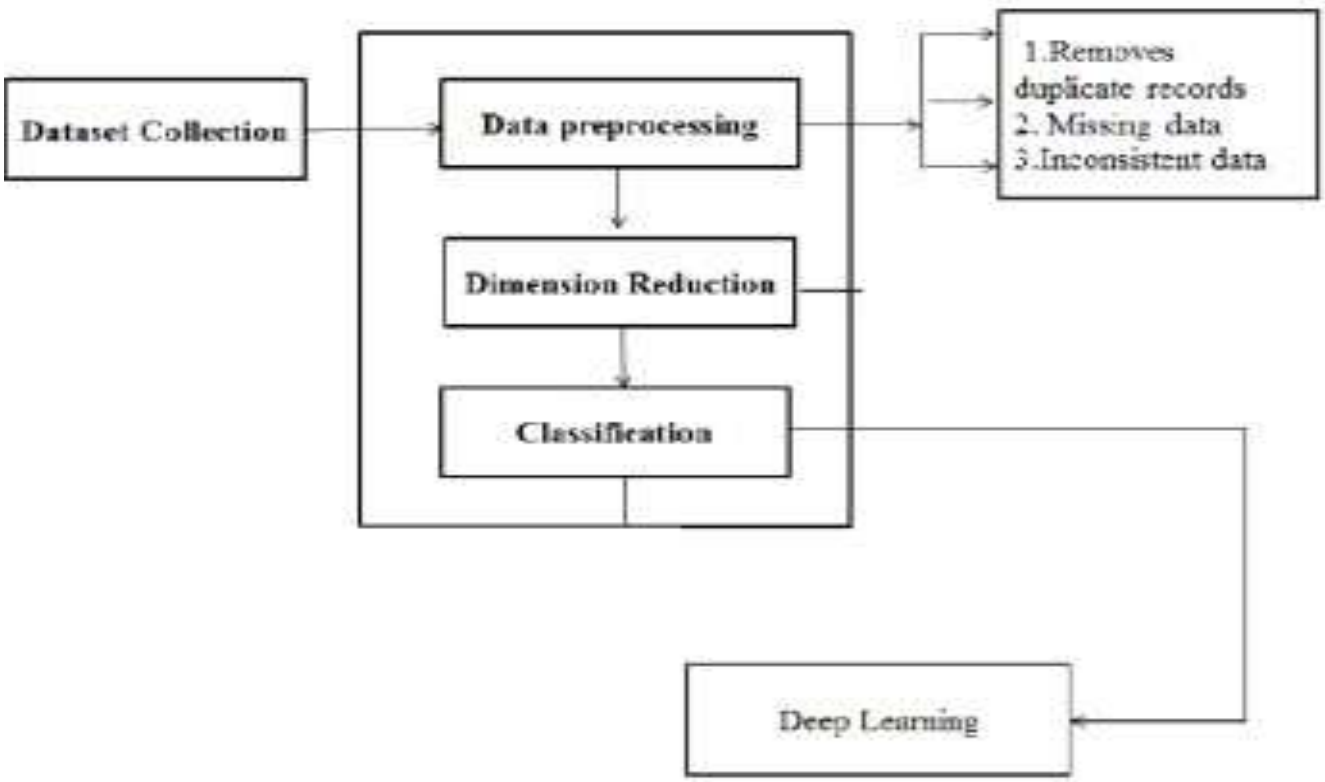
A data-flow diagram is a way of representing a flow of data through a process or a system. The DFD also provides information about the outputs and inputs of each entity and the process itself. A data-flow diagram has no control flow, there are no decision rules and no loops. Specific operations based on the data can be represented by a flowchart



D. ARCHITECTURE DIAGRAM

It is a pictorial way of conveying what the proposed system is intended to do. The first step is the data collection, where the data is collected from various sources which is further sent to data pre-processing. Once the data is pre-processed it can be Used for processing.

,



VII. IMPLEMENTATION

We use techniques of supervised learning in which the class Label of each training tuple is already known to develop a classifier model which predicts the categorical labels. The following are supervised learning algorithms which are employed to build a classifier for the given data set. The classifier model classifies the credit card transactions as either fraudulent or not.

A. HIDDEN MARKOV MODEL

Hidden Markov Model (HMM) is a statistical Markov model in which the system being modelled is assumed to be a markov process call it-X with unobservable ("hidden") states. HMM assumes that there is another Y process whose behaviour "Depends" on X the goal is to learn about X by Y observing HMM stipulates that, for each time instance, the conditional probability distribution of Y_{no} given the history $\{X_n = x_n\}$ $n<=n_0$ must not depend on $\{x_n\}$ $n<n_0$.

Hidden Markov models are known for their applications to thermodynamics, statistical mechanics, physics, chemistry, economics, finance signal processing, information theory, pattern recognition - such as speech, handwriting, gesture recognition, part-of-speech tagging, musical score following,[2] partial discharges and bioinformatics.

Markov Model: Series of (hidden) states
z= {z_1, z_2……} drawn from state alphabet
S= {s_1, s_2 ….s_|S|} where z_i belongs to S.

Hidden Markov Model: Series of observed outputx =
{x_1, x_2 …} drawn from an output alphabet V= {v1, v2. . . v_|v|} where x_i belongs to V.

Assumptions of HMM

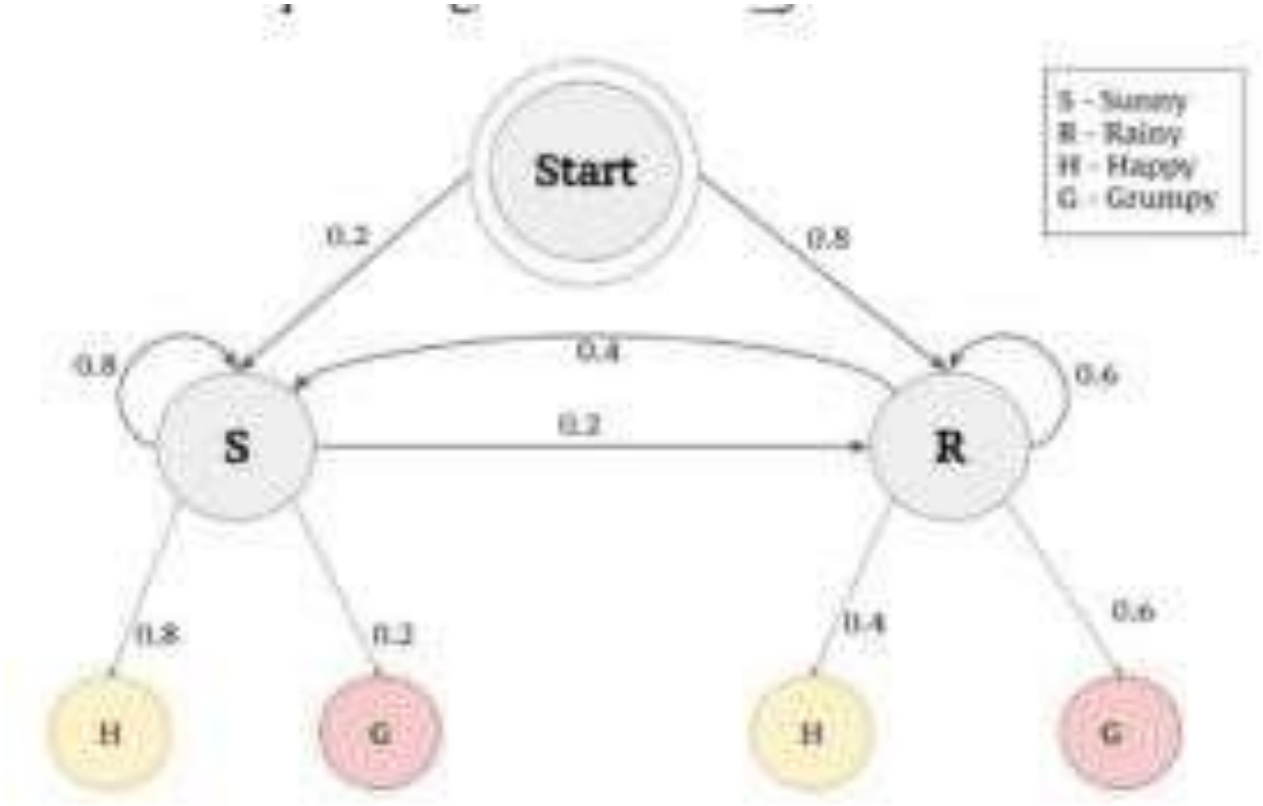
HMM too is built upon several assumptions and the following is vital.

Output independence assumption: Output observation is conditionally independent of all the hidden states and all other observations when given the current hidden state.

$$P(x_t=v_i / z_t=s_i) = P(x_t=v_i / x_1, x_2, .., x_T, z_1, z_2, .., z_T) = B_{ii}$$

Emission Probability Matrix: Probability of hidden state Generating output v_i given that state at the corresponding time was s_j.

$$\alpha\_i(t) = P(x\_1, x\_2, \ldots, x\_t, z\_t = s\_t; A, B)$$



Set of states (S) = {Happy, Grumpy}
Set of hidden states (Q) = {Sunny, Rainy}
State series over time = z∈ S_T
Observed States for four day = {z1=Happy, z2=
Grumpy, z3=Grumpy, z4=Happy}

The feeling that you understand from a person emoting is called the **observations** since you observe them. The weather that influences the feeling of a person is called the hidden state since you can't observe it.
Forward Procedure – Calculate the total probability of all the observations (from t_1) up to for testing. It train the model using the training set. It test the model using the testing set.
Backward Procedure -- Similarly calculate total probability output value of all the observations from final time (T) to t. It infers a function from labelled training data consisting of a set of training examples. In supervised learning, each example is a pair consisting of an input object and a desired

B. SUPPORT VECTOR MACHINE

The aim of a SVM is to fit a hyper plane between data points in space, i.e. support vectors, such that the samples are separated by the largest gap possible. Classification occurs by determining which side the gap new data points fall on. Typically, a binary linear classifier is used although there exist non-linear methods for SVM. In this paper. The SVM algorithm used is from "scikitlearn" v0.18.1 using Python2.7. This SVM implementation provides a means to apply cost-based balancing which will be briefly explored.

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyper plane. In other words, given labelled training data (supervised learning), the algorithm outputs an optimal hyper plane which categorizes new examples.
1. Set up the training data for model creation

2. Set up SVM's parameters

3. SVM Trainer
4. SVM Predictor

C. STEPS INVOLVED IN THE IMPLEMENTATION

1. IMPORT DATA

When running python programs, we need to use datasets for data analysis. Python has various modules which help us in importing the external data in various file formats to a python program. In this example we will see how to import data of various formats to a python program. The csv module enables us to read each of the row in the file using a comma as a delimiter.

2. DATA CLEANING

Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted. This data is usually not necessary or helpful when it comes to analysing data because it may hinder the process or provide inaccurate results.

3. DATA VISUALIZATION

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.in our project we use scatter diagram, histogram to visualization.

4. SPLITTING TRAIN TEST

Train/Test is a method to measure the accuracy of your model. It is called Train/Test because you split the data set into two sets: a training set and a testing set.80% for training, and 20%

5. SUPERVISED MACHINE LEARNING:

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances**.**

6. EVALUATION METRICS

Evaluation metrics are used to measure the quality of the statistical or machine learning model. Evaluating machine learning models or algorithms is essential for any project. There are many different types of evaluation metrics available to test a model. These include classification accuracy, logarithmic loss, confusion matrix, and others.

IX.CONCLUSION

In this paper, we have seen various methods that can be used to detect credit card fraud mechanism and also examine the result based on the principles of this algorithm. This method proves accurate in finding out the fraudulent transactions and minimizing the number of false alert.

The obtained results contradict results from prior research which suggest up sampling using techniques such as ADASYN increase performance in binary classification of highly imbalanced datasets. It is clear that when compared to conventional methods for dealing with class imbalance such as cost-based methods, or even under sampling, creating synthetic samples can result in much worse performance. In many cases it can actually handicap the classifier and produce results worse than not accounting for class imbalance at all. It is recommended that further research is done to examine the conditions upon which up sampling techniques such as ADASYN may successfully increase performance in extremely unbalanced datasets. Particularly it is suggested that a closer examination of the up sampled data in this paper is done to visualize and understand the characteristics of the synthetic samples and to evaluate to what degree they are representative of actual fraudulent samples.

Second, it is recommended that examination of the results of using ADASYN to up sample the minority class to a lesser degree is done, that is, rather than up sampling to produce equivalent class sizes, only up sample to reduce the class imbalance.
When the class imbalance is 99.8% biased towards the majority class, is feasible that attempting to up sample too much produces a dataset that when trained on a classifier, will result in a classifier that is always biased towards the minority class, indicated by high FP rates. Third, it is recommended that up sampling techniques such as ADASYN combined with conventional class imbalance mitigation techniques such as class reweighting are further explored. Although in this paper combining the two techniques lead to poorer performance compared to class reweighting alone, it is conceivable that by correcting for the performance impacts that ADASYN produced on the results in this paper, applying an additional technique that is shown to increase performance will result in a globally optimal FDS which may use off-the-shelf implementations of conventional classifiers.

REFERENCES

[1]. Investigating Hidden Markov Models Capabilities in Anomaly Detection,Author Shrijit S. Joshi and Vir V. Phoha,March 2005

[2]. Efficient anomaly detection by modeling privilege flows using hidden Markov model,Author Shinchondong, Sudaemoon-ku, Seoul 120-749, South Korea,14 February 2003.

[3]. Minority Report in Fraud Detection: Classification of Skewed Data, Author: C. Phua, D. Alahakoon, and V. Lee,june 2004.

[4]. Agent-Based Distributed Learning Applied to Fraud Detection, Author:S.Stolfo,A.L. Prodromidis ,April 21, 2011.

[5]. Neural Classifier with Fraud Density Map for Effective Credit Card Fraud Detection, AUTHOR: M.J. Kim and T.S. Kim, 20 August 2002.

[6]. An Experimental Study With Imbalanced Classification Approaches for Credit Card Fraud Detection, Author: Sara makki, zainab assaghir,yehia taher, rafiqul haque, mohandsaïd hacid, and hassan zeineddine,July 8, 2019.

[7]. Distributed Data Mining in Credit card Fraud Detection, Author Philip K. Chan, Florida Institute of Technology Wei Fan, Andreas
1. Prodromidir, and Salvotore 1. Stalfo, Columbia University, 1999.

[8]. Anonymous Credit Cards and Their Collusion Analysis, Author: Steven H. Low, Member, IEEE, Nicholas F. Maxemchuk, Fellow, IEEE, and Sanjoy Paul, Member, IEEE,6 dec 1996.

[9]. Neural Fraud Detection in Credit Card Operations, Author:Jos´e R. Dorronsoro, Francisco Ginel, Carmen S´anchez, and Carlos Santa Cruz, 4 july 1997.

[10]. Adversarial Learning in Credit Card Fraud Detection,Author Mary Frances Zeager, Aksheetha Sridhar, Nathan Fogal, Stephen Adams,Donald E. Brown, and Peter A.Beling,2017.

[11]. Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning, Author: Strate Andrea Dal Pozzolo, Giacomo Boracchi, Olivier Caelen, Cesare Alippi, Fellow, IEEE,and Gianluca Bontempi, Senior Member, IEEE,2018.

[12]. Real Time Credit Card Fraud Detection using Computational Intelligence, Author: Jon T. S. Quah and M. Sriganesh , 2007.

[13]. Credit card fraud detection using Machine Learning Techniques:A Comparative Analysis, Author :John O. Awoyemi, Adebayo O.

[14]. R. Vinoth, L. J. Deborah, P. Vijayakumar and N. Kumar, "Secure Multifactor Authenticated Key Agreement Scheme for Industrial IoT," in IEEE Internet of Things Journal, vol. 8, no. 5, pp. 3801-3811, 1 March1, 2021, doi: 10.1109/JIOT.2020.3024703.