International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-5, March 2020

Consumer Credit Risk Analysis using Data Mining Clustering and Business Intelligence Solutions

Subhash Babu Bathala, Muthuluru Nagendra

Abstract: ***In the recent years, the scale of online transaction has increased considerably. Subsequently, this has also increased the number of fraud cases, causing billions of dollars losses each year worldwide. Therefore, it has become mandatory to implement mechanisms that are able to assist in fraud detection. In this work, the use of Ensemble Genetic Algorithm is proposed to identify frauds in electronic transactions, more specifically in online credit card operations. A case study, using the dataset containing transactions made by credit cards in September 2013 by European cardholders, is used. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The presented algorithm achieves good performance in fraud detection as compared to the other machine learning algorithms. The results show that the proposed algorithm achieved good classification effectiveness in all tested instances.***

Keywords: Data mining, Credit risk analysis, ROC,

I. INTRODUCTION

Money has become the most significant requisite in every person's life and so the financial institutions in the current scenario, which create a cascade of financial datasets. The craze for e-commerce increased drastically over the years due to the appealing offers on online products. Credit cards are widely used for online payments, growing the trade volume in India and concurrently escalating the proportion of credit card fraud. The increase in the use of credit cards, in turn, increased the risks relevant to it. Risk refers to the loss of valuable information and in case of monetary theft; huge effect on the customers degrading the goodwill of the credit card issuer for their vulnerable data storage methods is possible. Hence, credit card fraud detection and prevention become the prime focus of risk control units of the Financial Institutions. Data mining (DM) is an efficient process for finding frauds that are found to be profitable for analyzing datasets. The extracted knowledge or hidden patterns are helpful in detecting fraudulent actions and for risk analysis in the future.

Revised Manuscript Received on March 30, 2020.
***** Correspondence Author
Mr. Subhash Babu Bathala*, Research Scholar, Department of Computer Science and Technology, Sri Krishnadevaraya University, Ananthapur, Andhra Pradesh, India.
Dr. Muthuluru Nagendra, Professor, Department of Computer Science and Technology, Sri Krishnadevaraya University, Ananthapur, Andhra Pradesh, India.

Credit card fraud is a crucial problem in today's financial markets. Credit card fraud detection is not a straight forward task mainly because of two reasons:

- the fraudulent behaviours usually differ for each attempt and
- The dataset is highly imbalanced, i.e., the frequency of majority samples (genuine cases) outnumbers the minority samples (fraudulent cases).

To deal with this problem, many researchers have focussed on detecting fraudulent behaviours using advanced machine learning techniques. This work proposes the notion of classification and the main strands of research in this area. It gives an overview of various machine learning algorithms and their comparison with Ensemble Genetic algorithm. The performance of each model is evaluated based on accuracy, recall, precision, f1 score, precision-recall (PR) curve and receiver operating characteristics (ROC) curve. The experimental results showed that the Ensemble Genetic Algorithm (Genetic model in combination with neural network X Gradient boost) performed better than other models [1].

In order to tackle this problem, data-level approach, where different resembling methods such as under sampling and oversampling strategies based on SMOTE have been implemented along with an algorithmic approach where ensemble models such as bagging and boosting have been applied to a highly skewed dataset containing 284807 transactions. Out of these transactions, only 492 transactions are labelled as fraudulent. Predictive models such as logistic regression, random forest, and XGBoost in combination with different resembling techniques have been applied to predict if a transaction is fraudulent or genuine. The remainder of this article has been organised as stated below. Some recent related works presents the proposed methodology and applies the proposed approach to real fraud detection on PCA dataset. Finally, gives the conclusions and directions for future improvement. The financial sector *is* an agency that has a significant job in the development of the economy of the nation. The obscure future practices of the clients are very critical to Customer Relationship Management (CRM). It turns out to be progressively significant for the bank to anticipate their customer's future choices so as to take appropriate activities in time. There are different zones where information mining and AI can be utilized in monetary divisions like client division and benefit, credit investigation, anticipating installment default, showcasing, false transactions, positioning speculations, improving stock portfolios, money the board and determining activities, high hazard advance candidates, most gainful MasterCard client and strategically pitching. In 2010, M. C. Lee and C.

To depicted the utilization of novel information-digging procedures for assessment of the endeavor money related pain and credit expectation; there are improved the presentation of algorithms by utilizing Support Vector Machine (SVM) with 3-folds cross approval and Back Propagation Neural Network (BPN) by the four estimated characteristics. The information for this examination has been gathered from the database of a security firm in Taiwan. Right now, are utilized test tests for preparing information and tests for testing information. By looking at the outcomes, there has been demonstrated that SVM gives higher exactness of about 100% expectation precision and grouping precision, suggesting low mistake rates, while BPN has prompted 96% of forecast exactness and 95% of characterization exactness [2].

Numerous sorts of research about client credit strategy examination were acted in 2012. K. Chopde et al. have examined the information-digging procedures for credit hazard investigation - specifically, the choice tree methods. This exploration utilized information digging for credit chance examination empowering the bank to diminish manual mistakes. This basic leadership pro¬cess is quick, it spares time handling and it encourages the bank to diminish the misinterpretations. The exploration result found by the Meta Decision Tree (MDTs) utilized a base level classifier and the Random Forest (RF) classifier, prompting a more precise characterization score than the CART choice tree. Generally speaking, the choice tree has end up being a method that can characterize the clients straightforwardly with a decent score and in this manner it can lessen the misfortune for the money related foundations in the most ideal manner [3].

I. G. Ngurah et al used to recommend a choice tree model for credit appraisal. This paper plans to recognize factors that are vital for a country bank in Bali to evaluate credit applications. Current choice criteria in credit chance appraisal are assessed. The credit hazard appraisal model has been applied to PT BPR X and it has utilized C5.0 philosophy; this model has utilized 84% of 1028 information as assessment information to recommend the new criteria in investigating the advance application. The outcome indicated that PT BPR X can diminish nonperforming advances to under 5% and the bank can be characterized or not as a well-performing one by applying information mining innovation. Around the same time, W. Chen et al proposed a half and half information mining procedure to manufacture a precise credit scoring model to assess credit chance dependent on the credit informational collection gave by a nearby bank in China.

This examination has proposed two preparing stages: the principal (bunching stage), implying that the examples of acknowledged and new candidates are gathered into a homogeneous cluster by utilizing K-implies grouping. The subsequent preparing stage is the grouping with Support Vector Machines (SVM). By examination with other credit scoring models, here the examples the past model uses three or four classes instead of two (great and awful credit). In addition, information mining thoughts and calculations can be applied to board information to discover the information that is unique in relation to the relapse information found by the customary straight relapse. Hence, G. Nie et al. proposed separation estimation with genuine board information about charge card application in China; it very well may be utilized in board information grouping with the K-implies bunching strategy. This examination dissected the gatherings of various clients by the behavior of credit cardholders. The outcome has indicated that progressively precise information can be found with the board information structure; separation estimation can mirror the information of various periods and board information can be utilized in groups to give new information [4].

II. RELATED WORK

In this section, some recent related works are discussed.

Bolton and Hand (2001) [1] proposed a technique of unsupervised detection which was based on breakpoint analysis. The analysis was used to identify changes in spending pattern. The advantage of breakpoint analysis is that balanced data is not required. However, the approach was in early development stage and the results which were presented were restricted to simple examples.

In Chen, Luo, Liang and Lee (2005) [2] employed Support Vector Machines (SVM) and Artificial Neural Networks (ANN) to investigate the time-varying fraud problem. The results exhibited that ANN is better than SVM in terms of training accuracy. However, SVM is better than ANN for predicting future data.

Rongchang Chen et al. in 2005 suggested a novel questionnaire-responder transaction (QRT) approach with SVM for credit card fraud detection. The objective of this research was the usage of SVM as well as other approaches such as Over-sampling and majority voting for investigating the prediction accuracy of their method in fraud detection. The experimental results indicated that the QRT approach has high degree of efficiency in terms of prediction accuracy [5].

Tung-shou Chen et al. (2006) proposed a binary support vector system (BSVS), in which support vectors were selected by means of the Genetic Algorithms (GA). In proposed model self-organizing map (SOM) was first applied to obtain a high true negative rate and BSVS was then used to better train the data according their distribution.

Guo (2008) used a confidence-based neural network to model a sequence of operations in credit card transaction. To ensure the accuracy of the fraud detection mechanism, the ROC analysis (Receiver operating characteristic) was introduced. In the model, initially, the confidence-based neural network was trained with synthetic data. If the incoming credit card transaction was found to be not accepted by the trained neural network model (NNM) with sufficiently low confidence, it was considered to be a fraud case.

Qibei Lu et al. in 2011 established a credit card fraud detection model based on Class Weighted SVM. Employing Principal Component Analysis (PCA), they initially reduced data dimension to less synthetic composite features due to the high dimensionality of data.

Y. Shahin and E. Duman in 2011 proposed fraud detection in credit card using a combination of Support Vector Machines (SVM) and Decision Trees. Decision trees outperformed SVM when the size of dataset was small but with increase in size of dataset, SVM reached accuracy of Decision trees. Andrea Dal Pozzolo et al. at Universite Libre de Bruxelles (2015) conducted a research on credit card fraud detection using Random Forest algorithm and proved it to be the best approach in the fraud detection task [4].Ishu Trivedi et al, (2016) presented a model of credit card fraud detection based on the principles of Genetic algorithm. The main aim of his research was to detect the fraudulent transactions and to develop a method of generating test data.

Geoffrey F. Miller, Peter M. Todd and Sailesh Hegde have elaborated the concept of designing of Neural Networks using Genetic Algorithms. They have stated the problem associated with intuitive network design by humans a proposed the idea of an automated evolutionary design method based on genetic algorithms as a solution to it.

Awoyemi et al. (2017) have performed a data analysis on credit card transaction data using K- nearest neighbour, logistic regression and Naïve Bayes. They concluded that K-nearest neighbour over took the other two.

Heta Naik and Prashasti Kanikar (2019) performed a comparative analysis of machine learning algorithms in detection credit ard fraudulations and concluded that Logistic regression and Ada Boost algorithms performed better in fraud detection.

After comparing the above cited papers, it is concluded that the authors were not able to find a fraud detection algorithm that addresses and combines all the features presented in this work.

The developed algorithm is based on Genetic algorithm in combination with Neural network and XG Boost.

III. PROPOSED SYSTEM

Data Mining Techniques & Predictive Analysis

Predictive analytics, pattern recognition, and classification problems are not new. Since quite a while ago utilized in the money related administrations and protection enterprises, a prescient examination is tied in with utilizing insights, information mining, and game hypothesis to break down present and verifiable actualities so as to make expectations about future occasions.

Regression analysis: Regression models are the backbone of prescient investigation. The straight relapse model breaks down the connection between the reaction or ward variable and a lot of autonomous or indicator factors. That relationship is communicated as a condition that predicts the reaction variable as a direct capacity of the parameters.

Choice modeling: Choice displaying is a precise and universally useful instrument for settling on probabilistic expectations about basic leadership conduct. It benefits each association to focus on its promoting endeavors at clients who have the most noteworthy probabilities of procurement. Decision models are utilized to recognize the most significant factors in driving client decisions. Ordinarily, the decision model empowers a firm to register a person's probability of procurement, or other social reaction, in view of factors that the firm has in its database, for example, geo-socioeconomics, past buy conduct for comparative items, frames of mind, or psychographics.

Rule induction: Rule acceptance includes creating formal guidelines that are extricated from a lot of perceptions. The guidelines removed may speak to a logical model of the information or neighborhood designs in the information. One significant principle enlistment worldview is the affiliation rule. Affiliation rules are tied in with finding fascinating connections between factors with regard to enormous databases. It is a method applied in information mining and uses rules to find regularities between items. For instance, on the off chance that somebody purchases nutty spread and jam, the individual in question is probably going to purchase bread. The thought behind affiliation rules is to comprehend when a client does X, the individual will undoubtedly do Y. Understanding those sorts of connections can help with gauging deals, limited-time valuing, or item arrangements.

Network/Link Analysis: This is another method for a partner like records. Connection examination is a subset of system investigation. It investigates connections and relationships among numerous objects of various sorts that are not evident from confined snippets of data. It is normally utilized for extortion location and by law requirement. You might be acquainted with the connecting investigation since a few Web-search positioning calculations utilize the system [7].

Clustering/Ensembles: Cluster analysis, or grouping, is an approach to arrange an assortment of "objects, for example, review respondents, into gatherings or bunches to search for designs. Outfit examination is a more current methodology that uses various bunch arrangements (a troupe of potential arrangements). There are different approaches to bunch or make outfits. Notwithstanding the strategy, the reason for existing is commonly the equivalent—to utilize bunch examination to segment into a gathering of sections and target markets to more readily comprehend and anticipate the practices and inclinations of the fragments. Grouping is a significant prescient examination approach with regards to item situating, new item improvement, use propensities, item necessities, and choosing test markets.

Neural networks: Neural networks Neural systems were intended to mirror how the mind learns and breaks down data. Associations create and apply fake neural systems to the prescient investigation so as to make a solitary structure. The thought is that a neural system is substantially more effective and exact in conditions where the complex prescient examination is required on the grounds that neural systems involve a progression of interconnected ascertaining hubs that are intended to delineate arrangement of contributions to at least one yield signals. Neural systems are perfect for getting significance from confused or loose information and can be utilized to extricate designs and recognize patterns that are too perplexing to possibly be seen by people or other PC methods. Advertising associations find neural systems helpful for foreseeing client requests and client division [6].

Decision Trees: Decision trees utilize genuine information mining calculations to help with grouping. A choice tree procedure will produce the standards followed in a procedure. Choice trees are valuable for helping you pick among a few strategies and empower you to investigate the potential results for different choices so as to evaluate the hazard and compensations for every potential game-plan. Such an examination is helpful when you have to pick among various methodologies or speculation openings, and particularly when you have restricted assets.

Financial Applications with Data Mining With Predictive Analysis:

Continuous prescient investigation implies you can extrapolate what has happened so far to foresee that something may be going to occur and forestall it. Think about a wild calculation. Under ordinary conditions, you can be observing the calculation's working parameters, which may incorporate what instruments are exchanged, size and recurrence of requests, request to-exchange proportion, and so on.

Constant checking implies moves can be made so as to affect the business. Utilizing client information, banks and other monetary establishments are applying the innovation to anticipate clients prone to stir and afterward making a move to keep the beat from happening. Prescient investigation distinguishes clients liable to agitate, at that point fragments those clients by productivity, volume, and length of commitment.

Predicting Abnormal Stock Market Returns: Insider merchants, for the most part, make strange returns as a result of the insider data accessible. Outcasts who can gain admittance to the insider data can likewise make expanded benefits. The capacity of untouchables, utilizing insider exchanging data, to anticipate irregular returns can be expanded by concentrating on information, for example, the size of the organization and the quantity of months later on that are prescient at stock costs. Right now, considers led by Safer, (2002) might be condensed as follows. The insider exchanging information utilized right now from January 1993 to mid-June 1997. The information was gathered from the Securities and Exchange Commission.

The stocks utilized in the examinations remembered all stocks for the S&P 600 (little top), S&P 400 (medium size top) and S&P 500 (enormous top) as of June 1997 that had insider records for the whole time of the investigation. There were 946 stocks in the three market tops which had accessible information in January 1993. From the rundown of 946 stocks, the example incorporated each stock that arrived at the midpoint of at any rate two purchases for every year. This brought about 343 stocks being utilized for the examination. The factors in the first informational index incorporate the organization, name, and rank of the insider, exchange date, stock value, number of offers exchanged to sort of exchange (purchase or sell), and number of offers held after the exchange. To survey an insider's earlier exchanging examples, the examination inspected the past nine and 18 weeks of exchanging history. The forecast time allotments for foreseeing anomalous returns were set up as three, six, nine, and a year. At that point, the information can be part into a preparation set (80% of the information) and approval set (20%). A neural system model is applied [8].

More secure found that the expectation of irregular returns could be upgraded in the accompanying manners:

- Extending the hour of things to come figure for upto one year.
- Increasing the time of back accumulated information;
- Narrowing the appraisal to specific ventures, for example, electronic hardware and business administrations and
- Focusing on little and moderate size instead of huge organizations.

IV. METHODOLOGY AND RESULTS

Data warehouses have a basic architecture that can create applications from data mining. Data mining can be considered as a result of the natural evolution of information technology from multiple disciplines as database and data warehouse technology, statistics, high performance computing, machine learning, computational intelligence (implying neural networks, fuzzy systems, evolutionary computing, swarm intelligence and so on), pattern recognition, data visualization, information retrieval, image processing, and spatial or temporal data analysis [9]. Data mining and Knowledge Discovery from the Database (KDD) are recent developments in the field of data management technologies9. KDD is a kind of data mining designed to extract knowledge from a large amount of data. The standard procedure in performing data mining based on Cross-Industry Standard Process of Data Mining (CRISP-DM) involves six phases.

These are the following:

- *Business understanding* phase, consisting in: choosing the objectives, understanding the business goal, learning situation assessment and developing a project plan.
- *Data understanding* phase, which consists of considering the data requirements and initial data collection, exploration and quality assessment.
- *Data preparation* phase, consisting in: selection of required data, data integration and formatting, data transformation and data cleaning.
- *Modeling* phase, consisting in: selection of appropriate modeling techniques, development and examination of alternative modeling algorithms and parameter settings and finding the tuning of model setting according to an initial assessment of the model's performance.
- *Evaluation* phase, corresponding to evaluation of the model experiment results.
- *Deployment* phase, representing an implementation step, where a model report is performed.

Clustering analysis are the information mining systems used to order as a factor or split into little gatherings of at least two. The items inside a gathering are like each other and not the same as the articles in different gatherings. K-implies grouping is a technique for bunching the observations into a particular number of disjoint clusters15,17,33. On a basic level, K-implies bunching plans to segment a dataset as {X1, X2,…, XN} into K subsets to limit the twisting measure characterized by the capacity given underneath

$$\sum_{n=1}^{n}\sum_{k=1}^{k} r_{nk} \left\| X_n - \mu_k \right\|^2$$

Where double pointer rnk=1, just if information point Xn is appointed to the kth cluster(for different cases, rnk =0) and μk signifies the mean of the kth cluster. We give an outline of the information structure and models and afterward present the aftereffects of model execution [10]. Data Mining Process Data mining, likewise prominently alluded to as KDD, is the extraction of examples speaking to information verifiably put away or caught in enormous databases, information distribution centers, the Web, other monstrous data stores or information streams. The engineering of a normal information mining framework may have the accompanying significant parts.

a. **Database, data warehouse, or other Information Repository:** This is one or a lot of databases, information distribution centers, spreadsheets, or different sorts of data archives. Information cleaning and information coordination systems might be performed on the information.

b. **Database or data warehouse server:** The database or information distribution center server is liable for getting the pertinent information, in view of the client's information mining demand.

c. **Knowledge base:** This is the area information that is utilized to direct the pursuit or assess the intriguing quality of coming about examples. Such information can incorporate idea progressions, used to arrange qualities or property estimations into various degrees of deliberation. Information, for example, client convictions, which can be utilized to survey an example's intriguing quality dependent on its suddenness, may likewise be incorporated. Different instances of space information are extra intriguing quality requirements or edges, and metadata.

d. **Data mining Engine:** This is basic to the information mining framework and in a perfect world comprises of a lot of utilitarian modules for undertakings, for example, portrayal, affiliation investigation, order, development, and deviation examination.

e. **Pattern Evaluation module:** This segment ordinarily utilizes intriguing quality measures and collaborates with the information mining modules in order to center the hunt towards fascinating examples. It might get to intriguing edges put away in the information base. On the other hand, the example assessment module might be coordinated with the mining module, contingent upon the usage of the information mining technique utilized. For proficient information mining, it is strongly prescribed to push the assessment of example intriguing quality as profound as conceivable into the mining procedure in order to discover the pursuit of just the fascinating examples.

f. **Graphical UI**: This module imparts among clients and the information mining framework, permitting the client to collaborate with the framework by indicating an information mining question or errand, giving data to help center the inquiry, and performing exploratory information mining dependent on the moderate information mining results.

What's more, this segment permits the client to peruse database and information stockroom mapping or information structures, assess mined examples, and imagine the examples in various structures. Information mining is iterative. Information mining process proceeds after an answer is conveyed. The exercises mastered during the procedure can trigger new business questions. Changing information can require new models. Resulting information mining forms profit by the encounters of past ones.

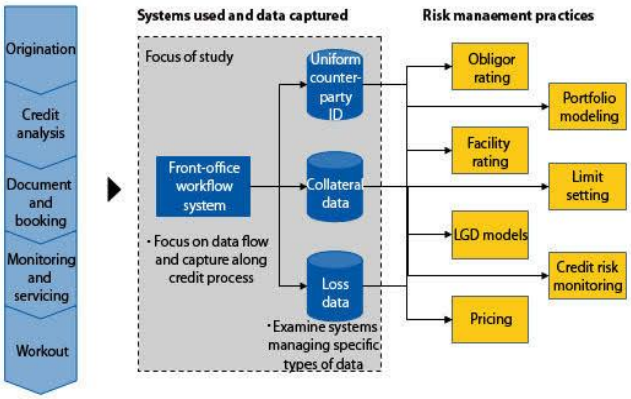

Figure.1: Risk Management Structure

The Data: The data set contains 117,019 lines, every one of them speaking to either default or not a default (double estimation) of a venture when they request an advance from a bank. Default and great wellbeing are described by the equivalent 235 marked factors that are legitimately acquired from the organizations: budget reports, monetary records, pay explanations and money flows= articulations where the qualities are considered at the most minimal degree of granularity. In the year 2016/2017, 115,288 lines spoke to organizations healthy and 1731 spoke to organizations in default. Due to the inclination made by imbalanced information, right now, give just outcomes adjusted preparing information of the twofold classes, following the technique reviewed in further steps. In the wake of bringing in the information, we cleaned the factors and evacuated highlights with no appropriate data (same incentive for every one of the endeavors; sign with no accessible passages like 'NaN' (Not a Number), for example) and we're left with 181 factors. At that point, we split the information into three subsets, thinking about 80% of the information (60% for the fitting and 20% for the cross-approval), and afterward, 20% of this information were utilized for test purposes. The approval execution licenses one to improve the preparation approach, and we use it to give expectation execution on the test set. In the preparation set, we confirm on the off chance that it is a reasonable dataset or not. Here it is: the estimation of zero speaks to 98.5% and the estimation of one 1.5%. In this way, extraordinary occasions are under 2%. Utilizing the SMOTE calculation depicted in Chawla et al. (2002), we get a decent set with 46% zero and 53% one [11].

The Models: The models we use have been nitty-gritty in the past area. We center around seven models: versatile net (strategic relapse with regularization), an arbitrary backwoods, a slope boosting displaying and a neural system approach with four unique complexities. To rank, the models concerning the organizations' financial soundness, the ROC bend and AUC criteria as RMSE criteria are utilized. An examination of the primary factors is given: first, we utilize the 181 factors (54 factors have been evacuated); at that point, we utilize the initial 10 factors chose by each model, Contrasting their presentation with deference with the models we use. An investigation of these factors finishes the examination. Figure below shows the flowchart of Genetic Algorithm:

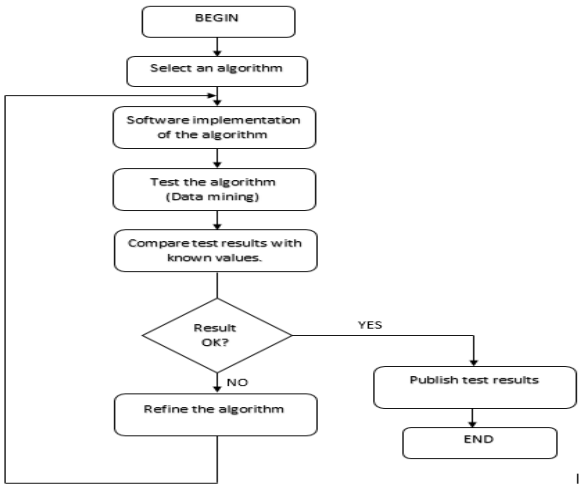

Figure.2 : Genetic Algorithm approach operation

The approach proposed in this paper is described along this section. It is composed of three main steps:

a. Data preparation:

i. **Data Standardization:** Data standardization is the foremost process of bringing the data into a common format before beginning collaborative research. It's important to standardize the features before applying the machine learning techniques. We performed standardization on both the "Time and Amount" feature using SMOTE algorithm.

Standardization can be achieved as follows:

$$z = \frac{x - \mu}{\sigma}$$

Where μ is the mean and σ is the standard deviation

ii. **Data splitting:** For each experiment, we split the entire dataset into 80% training set and 20% test set. We used the training set for resembling, hyper parameter tuning, and training the model and we used the test set to test the performance of the trained model. While splitting the data, we specified a random seed (any random number), which ensured the same data split every time the program executed.

iii. **Data resampling:** As already discovered that the dataset is highly unbalanced, the number of legitimate transactions outnumbers the number of fraudulent transactions. In this case, if we use this dataset to train our model, the model tends to be biased towards the legitimate transactions, and hence, it results in the poor performance of the model when tested in an unseen data.

To tackle this problem, we have used some resembling techniques such as random under sampling, random oversampling, SMOTE, Tomek links removal, a combination of SMOTE and Tomek links removal. We implemented these resampling techniques on the training data separately to make it balanced.

b. Ensemble Genetic Algorithm

In this step, we create a set of n elements of several neural networks by assigning hyper-parameters randomly to all Neural Networks, called population using fitness function based on the fitness score. The fittest neural network individuals called parents contribute to the population of next generation. Crossover rules combine two parents and mutation rules apply random changes to individual parents to form the next generation. In order to obtain better classification model, the rules at each iteration are subjected to reproduction, cross over and mutation. The process continues until the maximum number of generations is reached.

c. Data Prediction and Analysis

The final step of the predictive analysis is the performance evaluation of the model. In this thesis, we evaluated the performance of the models using confusion matrix, recall, precision, f1-score, precision-recall curve, and ROC curve [12].

The main parts of the proposed approach are described in the next subsections.

A: Ensemble Genetic algorithm

We have applied machine learning techniques to predict whether a credit card transaction is fraudulent or not. For this, we collected a publicly available dataset provided by the machine learning group of ULB (University Libre de Bruxelles), which contains the record of credit card transactions made by European cardholders and occurred in two days in September 2013. It contains 284,807 transactions out of which only 492 are fraudulent. The dataset is highly unbalanced as the positive class accounts for only 0.172% of the total transactions. When providing input data of a highly unbalanced class distribution to the predictive model, the model tends to be biased towards the majority samples.

As a result, it tends to misrepresent a fraudulent transaction as a genuine transaction. To tackle this problem, we implemented a data-level approach which includes various resampling techniques. In order to implement the machine learning algorithm on our huge imbalance dataset, we used SMOTE and sklearn module. Then, we analyzed all nine models with and without using resampling techniques. The comparison results revealed that the Genetic Algorithm in combination with Neural network and X Gradient Boost performed better than other models.

Figure.3: Risk Management analysis

The following outline summarizes how the genetic algorithm works:

- **Creating an initial population:** In this step, we create a set of n elements which is called a population. We begin the initialization by building random trees. The main parameters of this initialization method are the maximum number of nodes and the maximum depth. This is used to control the complexity of the solutions of the initial population. A function node is chosen at random that's acts as the root. At each iteration a terminal node is selected and included in the tree. The process stops when the maximum number of nodes is reached.
- **Defining a fitness function:** The fitness function determines how likely an individual is fit to be selected for reproduction, and this is based on its fitness score.

- **Selecting the parents:** The idea behind this step is to select the fittest individuals and let them pass their genes to the next generation. Two elements of the population are selected based on their fitness scores. The most commonly used selection methods include Roulette Wheel Selection, Rank Selection, Tournament Selection, Boltzmann Selection. We have employed the Roulette wheel selection method.

- **Roulette Wheel Selection:** Selection in this method is proportionate to the fitness of individual. Higher the fitness of individual, higher the chances of getting selected. The principle of roulette selection follows a linear search through a roulette wheel with the slots in the wheel weighted in proportion to the individual's fitness values. Roulette Wheel Selection is the easiest and simplest method to implement and consumes the least amount of time. However it suffers from problem of premature convergence which results in finding a solution which is locally optimum.

- **Making a crossover:** It is the most significant phase in a genetic algorithm. In this step, the individuals that remain after selection are subjected to crossover and mutations. A new population of *n* elements is reproduced from the selected elements by permuting and combining as many possible elements of the population.

- **Mutation:** After crossover, each individual can be mutated based on a given mutation probability. There are chances that from the crossover phase, we might get a population which will not contribute to the evolution of a new diverse population and our algorithm will converge prematurely. So we need to alter the sequence the elements from 1% of the newly created population to maintain this diversity. We can choose any sort of alteration.

The genetic algorithm stops when population converges towards the optimal solution.

Ensemble Genetic algorithm:

Define and initialize variables:
P*k*, *n*, Pf, Mr, *k* {Pop. size, no. of the individual in pop., Fraction of pop, Mutation rate, No. of generation}
set *k* = 0 {The initial generation}
Evaluate Pk:
compute fitness(*i*) for each *i* belongs P*k*;
do {
create generation *k* + 1;
//**crossover**
select Pf × *n* members of P*k*; pair them and produce offspring; insert offspring into P*k*+1;
//**mutation**
select Mr × n members of P*k*+1; perform random resetting on the randomly selected bit.
Evaluate P*k*+1:
Compute fitness (*i*) for each *i* belongs P*k*;
k = *k* + 1; // increment
}
While number of generation exceeded to defined limit return the fittest individual from P*k*.

We implemented ensemble genetic algorithm on our synthetic dataset that we got from the user through the web page and populating those data values to that respective dataset. The proposed algorithm tested all the values in the dataset and shows the current transaction is either fraud or not. On the other hand, in the console, it provides all the dataset transactions that are fraud and non-fraud. This helped us to know all the fraudulently of transactions in real-time.

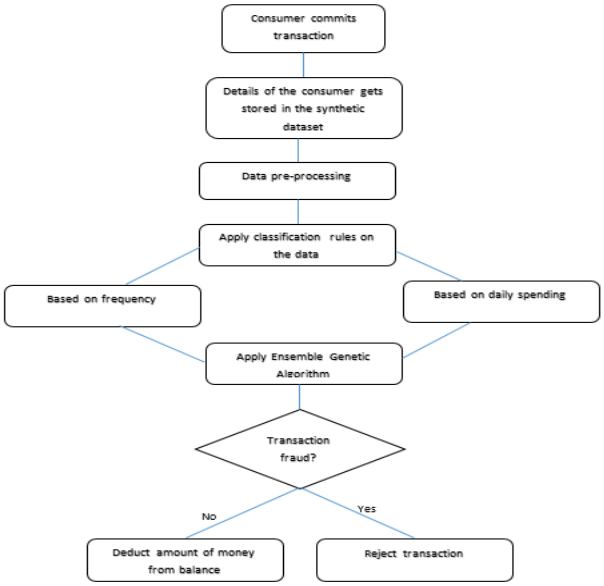

Figure.4: Flow chart of Operational Structures

The following figure shows the process of credit card fraud detection using our Ensembled model.
The figures below shows the evaluation metrics, the confusion matrix, the PR curve, and the ROC curve respectively of the Ensembled Genetic Algorithm. The model performed best in terms of precision value and accuracy.

	precision	recall	f1-score	support
0	0.87	1.00	0.93	79
1	1.00	0.76	0.86	50
accuracy			0.91	129
macro avg	0.93	0.88	0.90	129
weighted avg	0.92	0.91	0.90	129

Table.2: Classification Report

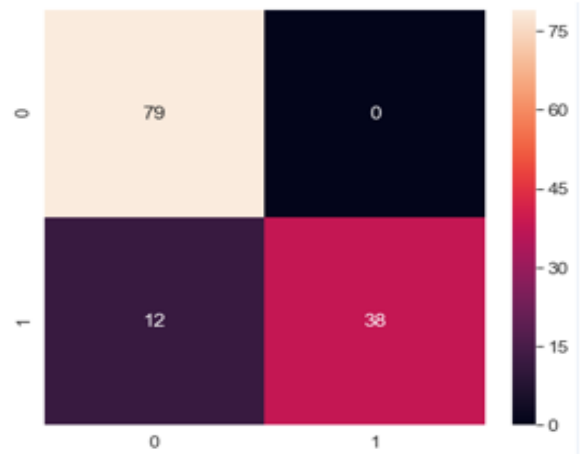

Figure.5 : Confusion Matrix-GA

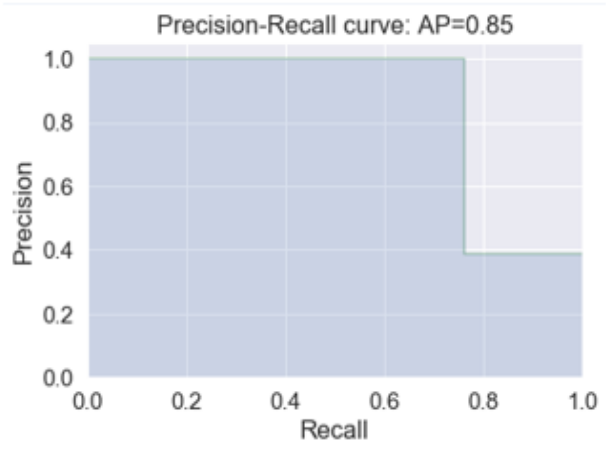

Figure.6 : Precision-Recall Curve - GA

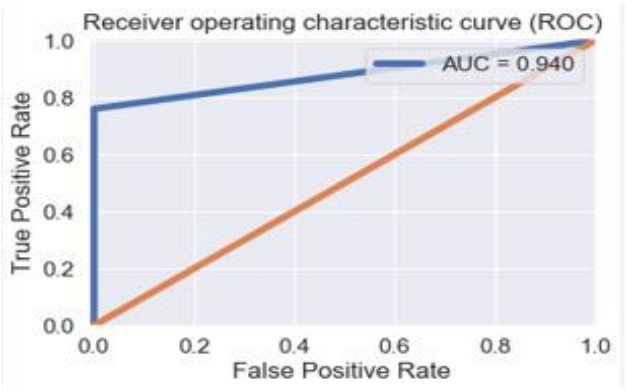

Figure. 7: ROC Curve-GA

V. EXPERIMENT – COMPARATIVE ANALYSIS

Table.3: Different parameters on result analysis

Models	Accuracy
Logistic Regression	0.90
Linear Discriminant Analysis	0.89
K – Nearest Neighbors	0.91
Decision tree	0.87
Support Vector Machines	0.91
X Gradient Boost	0.91
Random Forest	0.88
Gaussian Naive Bayes	0.90
Genetic Algorithm	**0.91**

Models	Precision
Logistic Regression	0.95
Linear Discriminant Analysis	1.00
K – Nearest Neighbors	1.00
Decision tree	0.87
Support Vector Machines	1.00
X Gradient Boost	1.00
Random Forest	0.90
Gaussian Naive Bayes	1.00
Genetic Algorithm	**1.00**

Models	Recall
Logistic Regression	0.78
Linear Discriminant Analysis	0.72
K – Nearest Neighbors	0.76
Decision tree	0.78
Support Vector Machines	0.76
X Gradient Boost	0.76
Random Forest	0.76
Gaussian Naive Bayes	0.74
Genetic Algorithm	**0.76**

Models	F1 score
Logistic Regression	0.86
Linear Discriminant Analysis	0.84
K – Nearest Neighbors	0.86
Decision tree	0.82
Support Vector Machines	0.86
X Gradient Boost	0.86
Random Forest	0.83
Gaussian Naive Bayes	0.85
Genetic Algorithm	**0.86**

A large number of possible alternatives require a joint assessment of several combinations in order to come up with a recommended approach. Since we are implementing the machine learning algorithm based on our huge imbalance dataset, we need to enhance the data in the dataset before processing it. We used SMOTE from sci-kit learn module to make it done.

We have compiled a comparison done on the KDD dataset on all of the techniques mentioned in previous section using four performance measures – Accuracy, Precision, Recall and F1 score**, s**ince these performance measures estimate the best model based on the rating on the scale from 0.0 to 1.0, the more it is closer to 1.0, the more it is better [10].

Table.4: Explanation of every Risk management module

Reference	Data mining tasks	Machine learning techniques	Research results
15	Clustering	K-means clustering	There can befound the panel data structure. There can reflect analysis for different periods.
29	Classification Clustering	SVM K-means clustering	There can be applied to panel data to find knowledge which is different from the regression knowledge discovered by the traditional linear regression.
34	Classification	SVM BPN	There are compared the results obtained by SVM and BPN for financial distress. The research results had shown that SVM leads to a lower error rate than BPN.
23	Classification	Decision trees – The CART model – MDT – RF	Decision trees techniques can reduce the manual errors, to obtainfaster and saving time processing,they can reduce the misjudgments, canclassify the customers directly and can reduce loss for the financial institutions.
24	Classification	Decision trees – PT BPR X, C.50	There is reduced the number of non-performing loans.
30	Classification	ANNS	There are classified the credit applications in order to allow the lenders taking a smart decision to select a loan application and to predictthe credit risk.
35	Classification Prediction	Decision trees – The CART model A survival analysis – Cox model – DA model – LR model	The presented results provide empirical evidence to support decision trees anda survival analysis in banks for financial distressto comparethe performance analysis. The CART model had obtained the best classificationaccuracy. In addition, the Cox, CART and DA modelshad led also to good prediction accuracy.
25	Classification	GLM Model GBM Model DRF Model	GBM model has shown a better performance. GBM had the highest probability of short- term recovery to support the activities of account managers and increase the efficiency of their approach with customers.

The above tables clearly show that Ensemble Genetic Algorithm performed better than the other machine learning models considering their overall f1 score. This shows the power of ensemble techniques that can give higher performance even in the presence of the class imbalance problem. Every model, when used with random under sampling, gave a good recall score but failed miserably in terms of precision [12].

VI. CONCLUSION

Even though there are many fraud detection techniques available today, but none is able to detect all frauds completely when they are actually happening; they usually detect it after the fraud has been committed. So we need a technology that can detect the fraudulent transaction when it is taking place so that it can be stopped then and there and that also in a minimum cost. So the foremost task is to build an accurate, precise and fast detecting fraud detection system for credit card frauds that can detect not only frauds happening over the internet like phishing, cross site scripting and site cloning but also tampering with the credit card itself i.e. it signals an alarm when the tampered credit card is being used. The proposed method has been extensively tested on different types of transactions. The results were promising, almost all the fraudulent transactions could be detected successfully and the proposed method has been compared with the existing method and the result shows that the proposed method performs better than existing methods. In this research fraudulent transactions have been detected and recognized which illustrates the robustness of the proposed system.

This proposed method enables the transaction at various types and improves the classification process, which can significantly improve the detection performance.

Credit card fraud is related to the non-stationary nature of transaction distributions in which the fraudsters usually always comes with a new way to attempt the fraudulent activities. Therefore, it becomes essential to consider these changing behaviors as well while developing a predictive model. Hence, a detailed study on dealing with non-stationary nature in credit card fraud detection can be performed. However, this study requires a huge amount of data.

Information mining dependent on AI systems is an innovation that can be utilized to investigate existing information, applications and client needs so as to manufacture and keep up long haul client connections. It can assemble certainty for customers making consumer loyalty and business the longest. Utilizing AI methods for classification and bunching assignments is famous in the advance installment forecast and the client credit strategy examination of the financial framework. Right now, proposed information mining strategies that contain two mains handling stages. The arrangement organize comprises a few models including SVM, ANNs, Decision Trees and BPN. We found that the SVM model and Decision Tree model are promising methods for an arrangement with money-related applications. The previously mentioned systems can diminish manual blunders, they can prompt quicker and sparing time preparing, they lessen them is decisions for grouping the clients legitimately and consequently, they can decrease the loss of the budgetary organizations.

In the grouping stage, K-implies bunching is the best performing model for client credit the executives of the credit scoring model. The scoring techniques are utilized to gauge the financial soundness candidate. At the point when credit advances and funds have the danger of being defaulted, credit chiefs need to create and apply information mining methods to deal with and break down credit information so as to spare time and lessen the mistakes. Information mining (actualized for the most part utilizing procedures of AI) will be a test for future research in banking and budgetary zones.

Beside future course is creating commonsense choice help programming apparatuses that make simpler to work in information-digging condition explicit for budgetary assignments, where hundreds and thousands of models, for example, neural systems, and choice trees should be investigated and balanced each day with another information stream coming each moment and checking the financial exchange. Inside the field of information mining in money, we expect a broad development of crossbreed techniques that join various models and give preferable execution over can be accomplished by people. In such an integrative methodology, singular models are deciphered as prepared fake "specialists". In this way, their blends can be composed also to the discussion of genuine human specialists.

Additionally, these fake specialists can be viably joined with genuine specialists. It is normal that these fake specialists will be worked as self-ruling keen programming operators. In this manner "specialists" to be joined can be information mining models, genuine monetary specialists, broker and virtual specialists that run exchanging rules separated from genuine specialists. A virtual master is a product insightful operator that is, generally, a specialist framework.

We instituted another term "master mining" as an umbrella term for removing information from genuine human specialists that is expected to populate virtual specialists. We additionally anticipate that the mixing with thoughts from the hypothesis of dynamic frameworks, tumult hypothesis, and material science of fund will extend.

REFERENCES

1. R. J. Bolton and D. J. Hand, “Unsupervised profiling methods for fraud detection,” Conference on Credit Scoring and Credit Control, September 2001.
2. R.-C. Chen, S.-T. Luo, X. Liang, and V. C. S. Lee, “Personalized approach based on svm and ann for detecting credit card fraud,” in Neural Networks and Brain, 2005. ICNN B ’05. International Conference on, vol. 2, 2005, pp. 810–815.
3. T. Guo, “Neural data mining for credit card fraud detection,” in International Conf. on Machine Learning and Cybernetics, vol. 7, July 2008.
4. Vapnik, Vladimir. 1995. The Nature of Statistical Learning Theory. Berlin: Springer.
5. Yitzhaki, Shlomo. 1983. On an extension of the gini inequality index. International Economic Review 24: 617–28.
6. Khandani, Amir E., Adlar J. Kim, and AndrewW. Lo. 2010. Consumer credit-risk models via machine-learning algorithms. Journal of Banking and Finance 34: 2767–87.
7. Huang, Zan, Hsinchun Chen, Chia-Jung Hsu,Wun-Hwa Chen, and SoushanWu. 2004. Credit rating analysis with support vector machines and neural networks: A market comparative study. Decision Support Systems 37: 543–58.
8. Gedeon, Tamás D. 1997. Data mining of inputs: Analyzing magnitude and functional measures. International Journal of Neural Systems 8: 209–17.
9. Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16: 321–57.
10. Bahrammirzaee, Arash 2010. A comparative survey of artificial intelligence applications in finance: Artificial neural networks, expert system and hybrid intelligent systems. Neural Computing and Applications 19: 1165–95.