DETECTING FRAUD IN CYBER BANKING USING FEATURE SELECTION AND GENETIC ALGORITHM

[1]G. Kavita,[2]CH. VijayaLaxmi, [3]D.Avanthika Shree , [4]Kavya Thati
[1]Asst prof, CBIT,[2]Asst prof, CBIT. [3]4th year 2nd sem CBIT. [4]4th year2nd sem CBIT.
[1]CSE Department,CBIT, Hyderabad, INDIA

Abstract : In the last decade,due to extensive development of information technology and communication infrastructure.there has been a rapid advancement in financial and banking system and Services. Banks and other financial institutions have invested in the field of modern technologies to provide more updated and efficient products and services. Thus, the variety of relevant products and services and also the number and value of transactions have increased. As online transactions became more and more popular, the frauds associated with them have also grown affecting the industry largely. Financial fraud has been a big concern for many organizations across industries, as billions of dollars are lost yearly because of this fraud. Securing transactions, detection of new ways of fraud and abuse in financial documents, the discovery of finished and unfinished frauds, detection and discovery of processes and operations of money laundering and etc. are among the most challenging issues in this area. The existing algorithms used do not give results considering different aspects of a transaction being carried out. However, there are a few researches which quote many features, but they are not practically implemented. Here a solution to the field of fraud detection in cyber banking is provided using feature selection and genetic algorithm. The bank data is given in an excel sheet and feature selection is applied to the data. To increase the accuracy of detected fraud, genetic algorithm is applied to the output of feature selection.

IndexTerms **- cyber banking, feature selection, genetic algorithm, fraud detection.**

I. INTRODUCTION

With the increase in the development of people's access to the internet, the use of online transactions in daily trades have increased. One of the most important problem of e-commerce is internet payment systems and fraud in e-payments.. Financial fraud can, not only cause financial damages to the relevant organization but also causes the loss of credit and damage to customer's confidence towards the system. Thus, in case of not using the fraud detection mechanisms, we should expect the increase of fraud statistics in e-banking system. Today, a large volume of financial and monetary transactions are performed on the internet. These services and transactions are not done in person. This makes the criminals remain unknown on the internet and encourages and stimulates the swindlers and fraudsters. Due to the lack of physical presence of customers in the context of electronic services, the need to recognize the identity for providing these services is very important and critical from the perspective of financial and monetary institutions. Perhaps it can be claimed that the main limitation in providing more extensive banking services is the need to recognize the identity of individuals. This issue is the most important factor of fraud attractiveness in the context of e-services and is increasing due to the development of e-banking services. Financial frauds can be widely classified as:

1. Bank fraud : It can be defined as "whoever knowingly executes to defraud a financial institution; or to obtain any of the money, funds, credits, assets, securities, or other property owned by a financial institution, by means of fraudulent pretends," that is, mortgage fraud, money laundering, etc.
2. Insurance fraud : It is the one which occurs in between the insurance process. It can happen while in application, billing, rating, claims, eligibility process etc. and are dedicated mostly by healthcare providers, consumers, agents or brokers, company employees and others.
3. Security and commodities fraud : It includes theft from manipulation of the market, securities accounts and wire fraud. It widely includes market manipulation, high yield investment fraud, commodities fraud, foreign exchange fraud, late-day trading, broker embezzlement, etc.
4. Other related financial fraud : It includes frauds such as mass marketing fraud and corporate fraud.

The fraud detection methods are divided into the two following main groups [17] :

1. Anomaly detection: In this method, the history of customer behaviour is considered a normal behaviour and any deviation from this behaviour can be recovered as an anomaly or fraud.
2. Misuse detection: This method focuses on specific behaviours of customer and assumes some unknown behaviours as a fraud. The main objective is to propose a new technique to detect fraud in e-banking using a new combination of algorithms to serve the purpose.

Financial fraud is normally discovered through outlier detection process enabled by data mining techniques, which also identify valuable information by revealing hidden trends, relationships, patterns found in a large database. Data mining, defined as "a process that uses statistical, mathematical, artificial intelligence, and machine learning techniques to extract and identify useful information and subsequently gain knowledge from a large database", is a major contributor for detecting different types of financial fraud through its diverse methods, such as, logistic regression, decision tree, support vector machine (SVM), neural network (NN) and naive Bayes.

II. REVIEW OF LITERATURE

The literature survey is the most important step in the software development process. Before building the system the below considerations and opinions are taken into account for developing the proposed system. The paper [1] using data from a credit card issuer, the neural network was trained on examples of fraud due to lost cards, stolen cards. It was possible to achieve a reduction of from 20% to 40% losses due to fraud.[2]This system predicts the probability of fraud on an account by comparing the current transactions and the previous activities of each holder. [3] The usability and applicability of the synthetic data is used, to train a fraud detection system. The system is then exposed to a set of authentic data to measure parameters such as detection capability and false alarm rate.In [4], Artificial Immune System for Fraud Detection (AISFD), system can perform online leaming with limited time and cost, and update the capability of fraud detection in the rapid growth of transactions and commerce activities due to the use of genetic algorithms.[5] It categorises, compares, and summarises relevant data mining-based fraud detection methods and techniques in published academic and industrial research. [6] applies Artificial Immune Systems(AIS) for credit card fraud detection and compares it to other methods such as Neural Networks and Bayesian Networks, Naive Bayes and Decision Trees. In paper [7] optimizing the parameters, using hybrid feature selection and genetic algorithm were shown.The paper [8] proposes a hybrid genetic rule learning algorithm by incorporating feature selection technique. When adding a relevant feature, the corresponding rule condition is also adjusted to improve the rule quality. In paper [9], a mode by using hybrid feature selection and anomaly detection algorithm in order to detect fraud in credit cards is presented. [10] presents a comparative study of five classification methods (Neural Nets(NN), Bayesian Nets(BN), Naive Bayes(NB), Artificial Immune Systems(AIS) and Decision Trees(DT)). The best models are obtained with DT, second best models are obtained with AIS. Next, BN and NN, follow equally, and finally, NB.In [11], the authors focused on 3 important issues of data imbalance, inconsistency, and assessment of methods. Experiments on large dataset of real-world transactions show that the alert precision, which is the primary concern of investigators, can be substantially improved.[12] states that feature selection algorithm in intrusion detection, data mining and pattern recognition plays a crucial role, it deletes unrelated and redundant features of the original data set to the optimal feature subset which are applied to some evaluation criteria.The study in [13] shows, winning the trust of customers through activities such as the safe processing and transmission of highly confidential information could be a useful step toward preserving electronic customers.The paper [14] presents review of various fraud detection techniques and discuss the issues regarding financial dataset used in fraud detection technique, which affects the accuracy of the fraud detection.In [15] by performing over sampling and extracting the principal direction of the data we can use KNN method to determine the anomaly of the target instance. Hence the KNN method can suit for detecting fraud with the limitation of memory. [16] Parallel big data framework with a machine learning algorithm is presented to handle offline massive transaction logs. This evaluation leads us to identify gaps and challenges to consider in the future.[17] Fraud was detected by using Genetic algorithm and whale algorithm.

III. PROPOSED METHOD

To find fraud in a bank transaction, following methods are used:
1. Feature Selection Algorithm
2. Genetic Algorithm

3.1 Fraud detection Environment

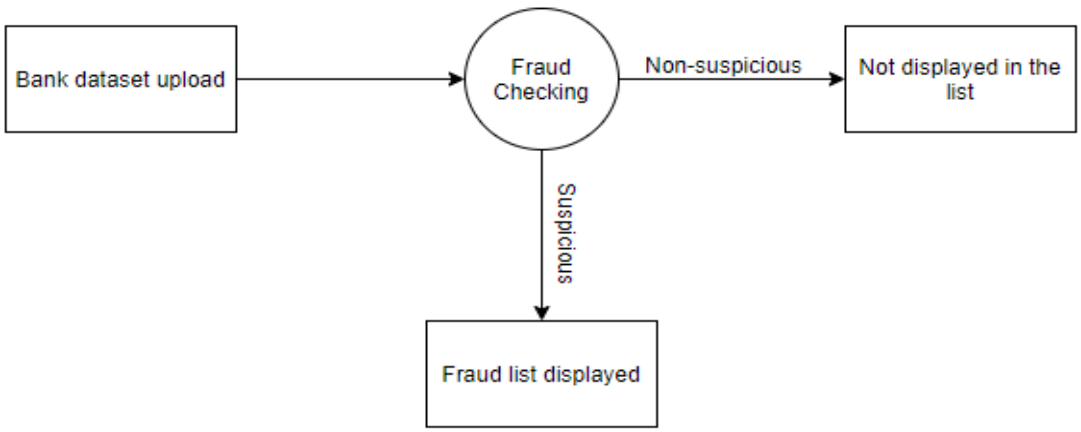


Figure1: Fraud Detection in E-banking-Data flow diagram –level0

The initiation of money transfer by the customer is given as input to the fraud checking process. It gives two results of non-suspicious(payment is made) and suspicious(re-checking is done and if still suspicious ,customer is notified and payment is not made).

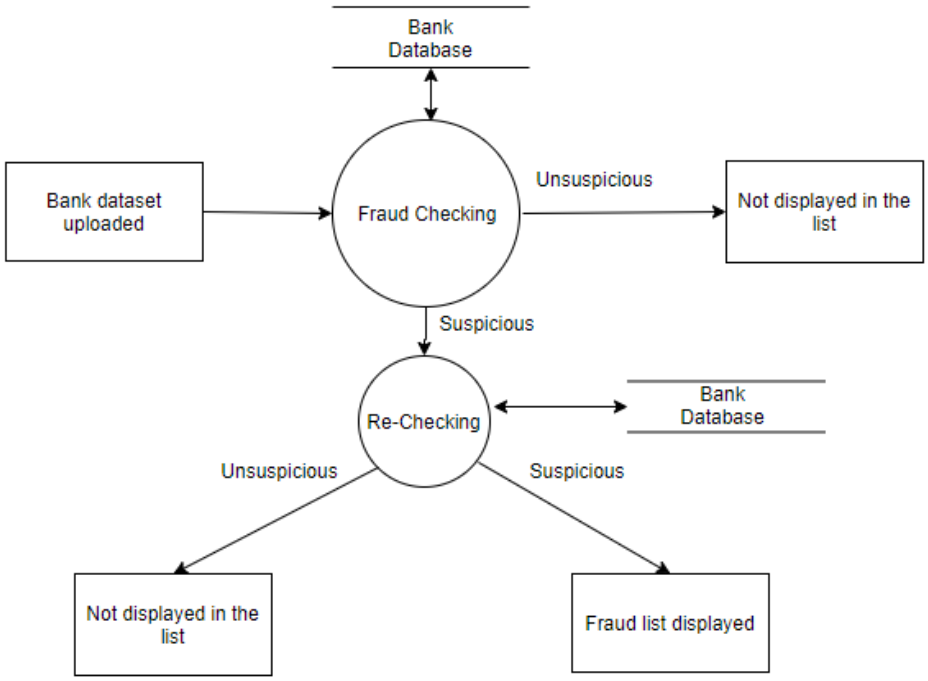


Figure2: Customer initiates money transfer and if suspicious, rechecking of fraud is done or else, payment is made. After rechecking, if there is any fraud in data notify the customer and defer the payment order or else make payment successfully.

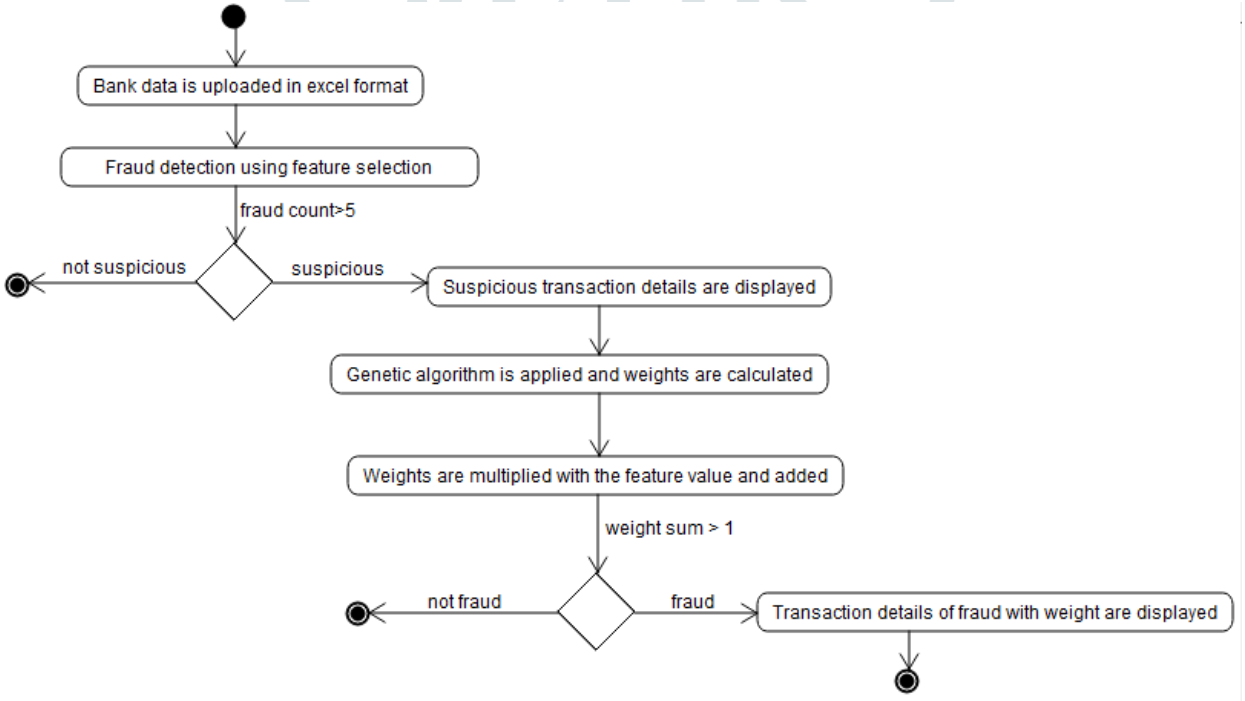


Figure 3 : *Fraud Detection in E-Banking* -The facts that are suspicious for money laundering are checked by taking the analytical information from the database. If the payment order turns out to be non-suspicious, then it is forwarded and the order is fulfilled.

IV. Feature Selection Algorithm:

In Machine Learning and Statistics, Feature Selection, also known as variable selection, or variable subset selection, is the process of selecting a subset of relevant features for use in model construction. Feature selection techniques are used for four reasons:
1. simplification of models to make them easier to interpret by researchers/users
2. shorter training times
3. to avoid the curse of dimensionality
4. enhanced generalization by reducing overfitting (formally, reduction of variance).
This is an exhaustive search of the space, and is computationally intractable for all but the smallest of feature sets. The choice of evaluation metric heavily influences the algorithm, and it is these evaluation metrics which distinguish between the three main categories of feature selection algorithms: wrappers, filters and embedded methods.

Our Usage : We have manually decided to use few of the metrics i.e., features to detect if a transaction is fraud. The threshold value for each of the metric and the formula used is as follows:
Flag : Fraud -1, Not Fraud - 0
Result variable :Sc
Exponent variable : e = 2.7

Feature 1: To calculate the location percentage to find any deviation in the location to which transaction is being made by the customer.

Sc = (1–percentage_location_count / 100)
if (Sc>0.5):Flag = 1
else :Flag = 0

Feature 2 : To calculate the difference between current transaction amount and maximum amount that was ever sent by the customer.
X=(current_transaction_amount–max_transaction_amount) * 25 /(Current_transaction_amount)
Sc=(1 / (1 + e –X))
if (Sc>0.5) :Flag = 1
else :Flag = 0

Feature 3: To calculate the difference between current transaction total and maximum amount that was ever sent by the customer in a week.
X=(current_transaction_total–max_transaction_total) * 25 /(current_transaction_amount*7)
Sc=(1 / (1 + e –X))
if(Sc>0.5) :Flag = 1
else :Flag = 0

Feature 4: To calculate the difference between the time at which current transaction is made and the average time when all other transactions that are been made previously.
Sc=1-(currentTime-AvgTime)/100
if(Sc>0.5) :Flag = 1
else :Flag = 0

Feature 5: To calculate the difference between the time at which current transaction is made and the average number of late night transactions previously made by the customer.
Sc=1-(currentTime-LateN8AvgTime)/100
if(Sc<0.5) :Flag = 1
else :Flag = 0

Feature 6: To calculate the difference between the time at which current transaction is made and the last transaction previously made by the customer.
X=1-(currentTime-LastTime)/100
Sc=(1 / (1 + e –X))
if(Sc>0.5) :Flag = 1
else :Flag = 0

Feature 7: To calculate the frequency of the type of payment (credit, debit or withdrawal) being made when compared to the previous transactions made by the customer.
Sc = (count / tot) * 100
Sc = 1 - Sc / 100
If(Sc>0.5) :Flag = 1
else :Flag = 0

Feature 8: To calculate the frequency of the type of transaction (domestic or overseas) being made when compared to the previous transactions made by the customer.
Sc = (1–transaction_percentage / 100)
if(Sc>0.5) :Flag = 1
else :Flag = 0

V. GENETIC ALGORITHM:

A genetic algorithm (GA) is a metaheuristic inspired by the process of natural selection that belongs to the larger class of evolutionary algorithms (EA). Genetic Algorithm (GA) is a search-based optimization technique based on the principles of Genetics and Natural Selection. It is frequently used to find optimal or near-optimal solutions to difficult problems which otherwise would take a lifetime to solve. It is frequently used to solve optimization problems, in research, and in machine learning. Genetic algorithms are commonly used to generate high-quality solutions to optimization and search problems by relying on bio-inspired operators such as mutation, crossover and selection.

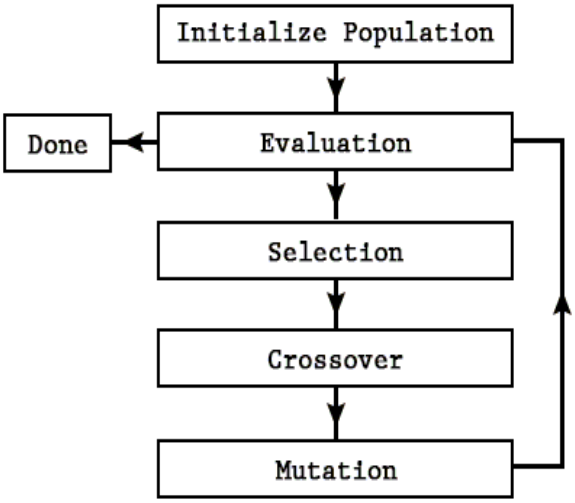


Figure 4 : *Genetic algorithm flow chart*

Inputs: Number of the weights (num_weights) we are looking to optimize, mating pool size (num_parents_mating), population size (sol_per_pop)

Output : Best solution, Best solution fitness.

Population Initialization: There are two primary methods to initialize a population in a GA. They are −

- Random Initialization − Populate the initial population with completely random solutions.
- Heuristic initialization − Populate the initial population using a known heuristic for the problem.

It has been observed that the entire population should not be initialized using a heuristic, as it can result in the population having similar solutions and very little diversity. It has been experimentally observed that the random solutions are the ones to drive the population to optimality.

Fitness function: Fitness Function (also known as the Evaluation Function) evaluates how close a given solution is to the optimum solution of the desired problem. It determines how fit a solution is to get an optimized solution. The fitness function calculates the sum of products between each input and its corresponding weight.

Selection: Selection is the stage of a genetic algorithm in which individual genomes are chosen from a population for later breeding (using the crossover operator). The best individuals in the current generation as parents for producing the offspring of the next generation are selected.

Crossover: The crossover operator is analogous to reproduction and biological crossover. In this more than one parent is selected and one or more off-springs are produced using the genetic material of the parents. Crossover is usually applied in a GA with a high probability. The point at which crossover takes place between two parents usually is at the center.

Mutation: mutation may be defined as a small random tweak in the chromosome, to get a new solution. It is used to maintain and introduce diversity in the genetic population and is usually applied with a low probability. If the probability is very high, the GA gets reduced to a random search. Mutation changes a single gene in each offspring randomly.

VI. RESULTS AND DISCUSSIONS

6.1 The Dataset:

We have generated the data of 10200 records from www.generatedata.com . 8000 records were used to train the model and testing was done on 2200 records.

It took 1.30 minutes to get the results on the data set of 2200 records, and 5 hours on the data set of 10000 records.

The software used is python ,MYSQL and SQLyog.

Schema :

transactionID (numeric) - Primary Key
CustomeraccountID (alphanumeric)
transactionAmount (numeric)
transactionDate (timestamp)
transactionTime (timestamp)
Customer_Bank (varchar)
Destination_accountID (alphanumeric)
Original Amount (numeric)
Payment_Type (varchar)
Destination_Bank (varchar)
transac_overseas (varchar)
City (varchar)

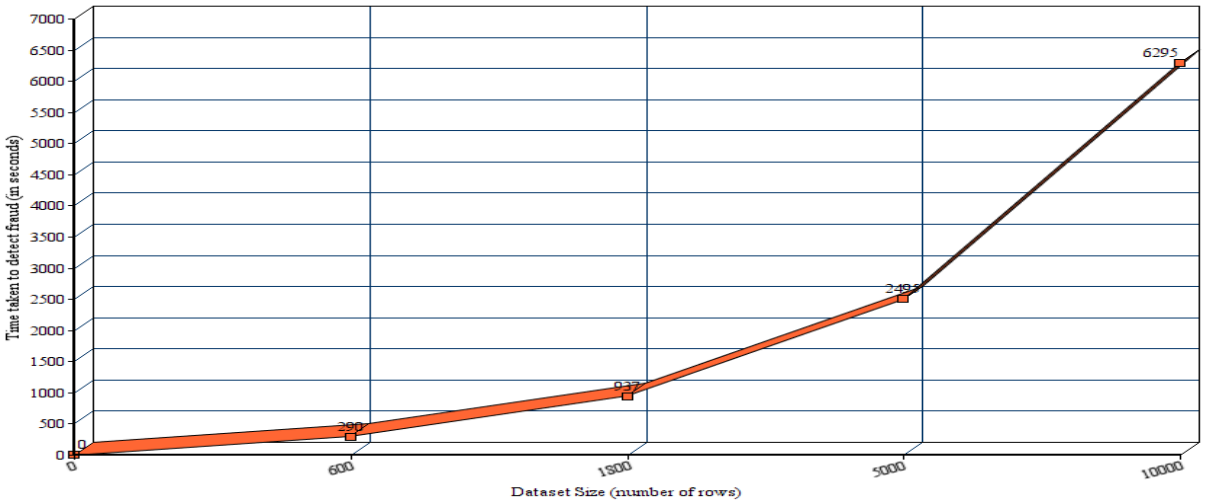
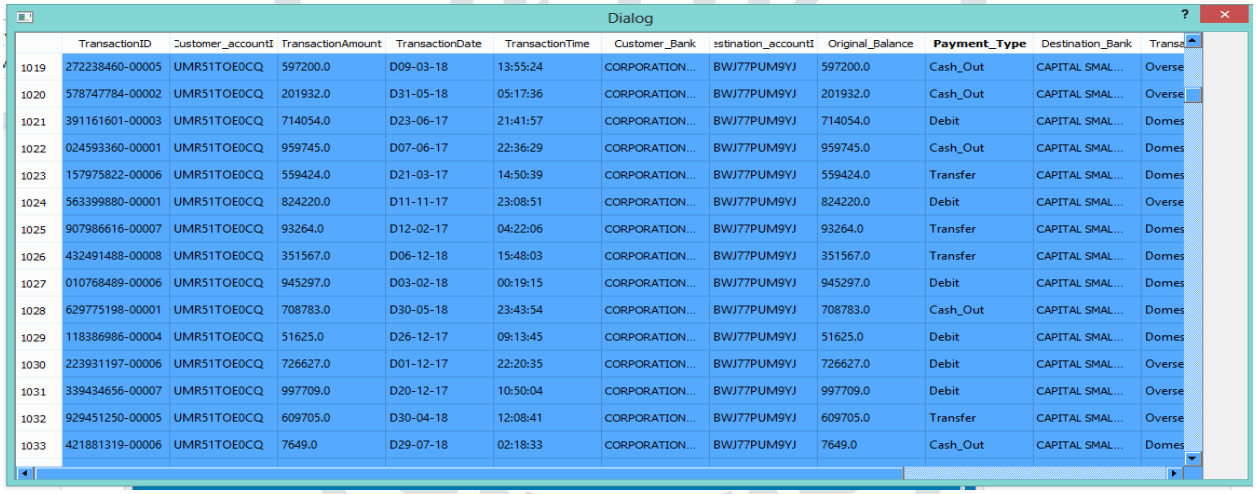


Figure 5 : *Relationship between the dataset size (in number of rows) and the time taken to detect fraudulent data in the dataset (in seconds)*

It is also observed that the upload time increases with the increase in the number of rows in the dataset.

6.2 Screenshots of the result from the model:



	TransactionID	Customer_accountI	TransactionAmount	TransactionDate	TransactionTime	Customer_Bank	estination_accountI	Original_Balance	Payment_Type	Destination_Bank	Transa
1019	272238460-00005	UMR51TOE0CQ	597200.0	D09-03-18	13:55:24	CORPORATION...	BWJ77PUM9YJ	597200.0	Cash_Out	CAPITAL SMAL...	Overse
1020	578747784-00002	UMR51TOE0CQ	201932.0	D31-05-18	05:17:36	CORPORATION...	BWJ77PUM9YJ	201932.0	Cash_Out	CAPITAL SMAL...	Overse
1021	391161601-00003	UMR51TOE0CQ	714054.0	D23-06-17	21:41:57	CORPORATION...	BWJ77PUM9YJ	714054.0	Debit	CAPITAL SMAL...	Domes
1022	024593360-00001	UMR51TOE0CQ	959745.0	D07-06-17	22:36:29	CORPORATION...	BWJ77PUM9YJ	959745.0	Cash_Out	CAPITAL SMAL...	Domes
1023	157975822-00006	UMR51TOE0CQ	559424.0	D21-03-17	14:50:39	CORPORATION...	BWJ77PUM9YJ	559424.0	Transfer	CAPITAL SMAL...	Domes
1024	563399880-00001	UMR51TOE0CQ	824220.0	D11-11-17	23:08:51	CORPORATION...	BWJ77PUM9YJ	824220.0	Debit	CAPITAL SMAL...	Overse
1025	907986616-00007	UMR51TOE0CQ	93264.0	D12-02-17	04:22:06	CORPORATION...	BWJ77PUM9YJ	93264.0	Transfer	CAPITAL SMAL...	Domes
1026	432491488-00008	UMR51TOE0CQ	351567.0	D06-12-18	15:48:03	CORPORATION...	BWJ77PUM9YJ	351567.0	Transfer	CAPITAL SMAL...	Domes
1027	010768489-00006	UMR51TOE0CQ	945297.0	D03-02-18	00:19:15	CORPORATION...	BWJ77PUM9YJ	945297.0	Debit	CAPITAL SMAL...	Domes
1028	629775198-00001	UMR51TOE0CQ	708783.0	D30-05-18	23:43:54	CORPORATION...	BWJ77PUM9YJ	708783.0	Cash_Out	CAPITAL SMAL...	Domes
1029	118386986-00004	UMR51TOE0CQ	51625.0	D26-12-17	09:13:45	CORPORATION...	BWJ77PUM9YJ	51625.0	Debit	CAPITAL SMAL...	Domes
1030	223931197-00006	UMR51TOE0CQ	726627.0	D01-12-17	22:20:35	CORPORATION...	BWJ77PUM9YJ	726627.0	Debit	CAPITAL SMAL...	Overse
1031	339434656-00007	UMR51TOE0CQ	997709.0	D20-12-17	10:50:04	CORPORATION...	BWJ77PUM9YJ	997709.0	Debit	CAPITAL SMAL...	Overse
1032	929451250-00005	UMR51TOE0CQ	609705.0	D30-04-18	12:08:41	CORPORATION...	BWJ77PUM9YJ	609705.0	Transfer	CAPITAL SMAL...	Overse
1033	421881319-00006	UMR51TOE0CQ	7649.0	D29-07-18	02:18:33	CORPORATION...	BWJ77PUM9YJ	7649.0	Cash_Out	CAPITAL SMAL...	Domes

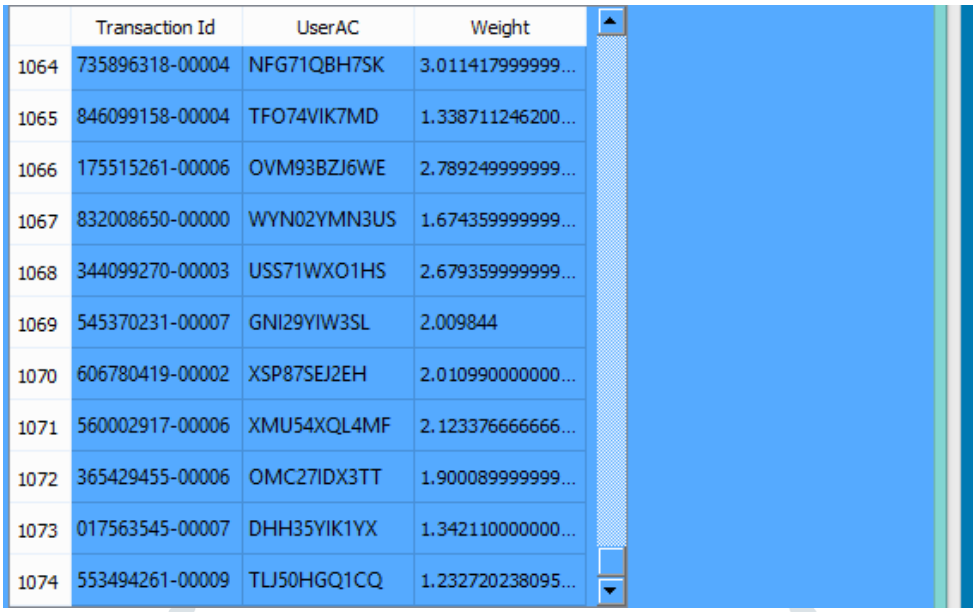
Figure 6 : *Uploaded Dataset*

This is the dataset we uploaded for detection. It contains TransactionID, Customer_accountID, TransactionAmount,Transaction_Date,Transaction_Time,Customer_bank,Destination_accountID, OriginalAmount,Payment_method,Destination_Bank,Transaction_overseas, Seconds,City.

	Transaction Id	UserAC	Fraud Count
2265	994292563-00008	TFO74VIK7MD	5
2266	994389401-00005	DHH35YIK1YX	5
2267	994431278-00005	SBS81RBR7JP	4
2268	996796306-00008	SBS81RBR7JP	4
2269	996852331-00007	GNI29YIW3SL	4
2270	997272885-00002	WYN02YMN3US	5
2271	997486857-00003	OMC27IDX3TT	4
2272	997682018-00004	LDX63PEX4GF	5
2273	998934624-00003	SBS81RBR7JP	5
2274	999072887-00006	DHH35YIK1YX	5
2275	999362734-00009	HHI08VQJ6DC	5

Figure 7: *Result of Feature Selection*

This page displays the transaction IDs that are having a fraud feature count more than 3.



	Transaction Id	UserAC	Weight
1064	735896318-00004	NFG71QBH7SK	3.011417999999...
1065	846099158-00004	TFO74VIK7MD	1.338711246200...
1066	175515261-00006	OVM93BZJ6WE	2.789249999999...
1067	832008650-00000	WYN02YMN3US	1.674359999999...
1068	344099270-00003	USS71WXO1HS	2.679359999999...
1069	545370231-00007	GNI29YIW3SL	2.009844
1070	606780419-00002	XSP87SEJ2EH	2.010990000000...
1071	560002917-00006	XMU54XQL4MF	2.123376666666...
1072	365429455-00006	OMC27IDX3TT	1.900089999999...
1073	017563545-00007	DHH35YIK1YX	1.342110000000...
1074	553494261-00009	TLJ50HGQ1CQ	1.232720238095...

Figure 8 : *Final Result Page*

This dialog gives the weight of each transaction from which we come to know if a particular transaction is completely fraud after passing it through the genetic algorithm.

VII. Conclusion

The proposed system is an efficient and effective method for fraud detection in cyber banking. It focuses on faster detection of a fraudulent transaction. The time taken to detect a transaction is fraud is very less when compared with other systems as the existing systems take longer time to train the model.

The efficiency of the model is also improved as we have used a combination of two algorithms where the results of one (feature selection) serve as input to other (genetic algorithm) for further analysis.

This also paves a path towards making an assumption that this system would be more accurate than the others.

References

[1] Credit Card Fraud Detection with a Neural-Network; Ghosh and D.L. Reilly; 1994.
[2] Detecting Payment Card Fraud With Neural Networks; Hassibi. K; 2000.
[3] Synthesizing Test Data for Fraud Detection Systems. Proc. of the 19th Annual Computer Security Applications Conference;Emilie Lundin Barse, Hakan KVarnstrom, Erland Jonsson; 2003.
[4] Artificial Immune System for Fraud Detection; Tue, Ren, Liu; 2004.
[5] A comprehensive survey of data mining-based fraud detection research; Clifton Phua, Vincent Lee, Kate Smith, and Ross Gayler; 2005.
[6] Credit Card Fraud Detection with Artificial Immune System; Manoel Fernando Alonso Gadi1, Xidi Wang3 and Alair Pereira do Lago; 2005.
[7] Comparison with Parametric Optimization in Credit Card Fraud Detection; Gadi, Wang, Lago; 2008.
[8] A Hybrid Genetic Algorithm for Simultaneous Feature Selection and Rule Learning Zhichun Wang, Minqiang Li; 2008.
[9] Data Mining Application for Cyber Credit-Card Fraud Detection System; Akhilomen. John; 2013.
[10] Case-based Reasoning; Kolodner. J, Morgan Kaufmann; 2014.
[11] Credit Card Fraud Detection and Concept-Drift Adaptation with Delayed Supervised Information; Pozzolo.A, Boracchi.G, Caelen.O, Alippi.C, Bontempi.G; 2015.
[12].A Hybrid Feature Selection Algorithm; Chunyong Yin, Luyu Ma, Lu Feng,JinWang,Zhichao Yin, Jeong-UK Kim; 2015.
[13] An Intelligent system for User Behavior detection in Internet Banking; Saeideh Alimolaei;2015.
[14] Online Fraud Detection Techniques: A Review on Data Mining Approaches; B.B.Sagar, Pratibha Singh, S. Mallika; 2016.
[15] Analysis on Credit Card Fraud Identification Techniques based on KNN and Outlier Detection; N.Malini, Dr.M.Pushpa; 2017.
[16] Fraud Risk Monitoring for E-banking Transactions; Guo,Wang, Dai, Cheng, Tongsen Wang; 2018.
[17] Fraud detection in E-banking by using the hybrid feature selection and evolutionary algorithms; Alireza Pouramirarsalani, Majid Khalilian, Alireza Nikravanshalmani; 2017.