Fraud Detection for Online Retail Using Random Forest

R. Abiramy, Kumar Narayanan, R. Anandan, C. Swaraj Paul

***Abstract:** The evolution of e-commerce has widely risked the electronic transaction over the past few years. This has significantly raised the issue of fake events worldwide with millions of buck deficits. This work aims to give a solution to frauds done through credit cards. Using datamining and machine learning techniques we provide a highly secured transaction to Web payment gateways (e.g. UPI).*

***General Terms---** Data Mining, Decision Tree, Random Forest, Bagging, Filtering Techniques.*

***Keywords---** Electronic Commerce, Credit Card Fraud, Fraud Detection, Online Banking Electronic.*

I. INTRODUCTION

The enormous development of e- commerce, smart devices and website purchase has become a daily means of purchase. This has significantly increased the growth of credit card users. This growth of payment system has also increased the credit card frauds. The processing of these datasets requires fast and efficient algorithms. Some of the fraud detection techniques are data mining, machine learning, decision theory, neural networks, etc. These methods being able to handle big data can resolve such challenges. Our research can be used to enhance performance in fraud detection and enhance security to prevent financial loss.

II. RELATED TERMS

Credit cards – are small laminated cards granted by the banks with pin number, allowing the cardholder to purchase goods on credit.

Credit card Fraud – this comprises identity theft, identity assumption and fraud sprees with the intension of avoiding payment.

Some real-world issues –

- Credit card fraud: 60 years old doctor duped of Rs 1.40 lakhs without any alerts or OTPs shared.
- Vadodara student arrested for booking tickets using the data of credit card fraud belonging to three US nationals.

Credit card Fraud Detection – frequently used fraud detection technique is data mining [4]. Data mining is mostly used to classify, cluster and segment the data. It can also naturally find associations and rules that express interesting patterns, including those related to fraud.

Data mining is widely used in retail, health care, credit card services, telecommunication, etc since it can deal with large data sets and reduce risk and more.

We have implemented random forest algorithm a data mining technique to enhance credit card fraud detection.

III. MATERIALS & METHODS

The proposed system mainly uses web services, the name itself indicates “services available over the WEB” e.g. java4s.com i.e., whenever we click a link in the browser it displays the content in HTML format. With the help of web services, we can interconnect various operations on several platform. The two sorts of web services are:

- REST (Representational State Transfer)
- SOAP (Simple Object Access Protocol)

We have used SOAP for our project. JAX-WS, Apache Axis2 supports the execution of SOAP. Software requirements include the following: Front end uses JSP and structs.

JSP is implemented using HTML and CSS. HTML is a standard mark-up language for creating web pages and web apps. Cascading Style Sheet is a mechanism of adding styles e.g. font, colour, spacing, etc.

Struts are Open Source Framework given by Apache software which is used to develop web application for java. Backend framework, AJAX, JavaScript are used in the server-side programming for web development and high performance.

MySQL database is a collection of data, stored and accessed from the computer system. The hardware requirements are:

Hard Disk	: 250GB and Above
RAM	: 4GB and Above
Processor	: I3 and Above (32 bit)

Manuscript published on 28 February 2019.
***** Correspondence Author (s)

R. Abiramy, Department of Computer Science & Engineering, Vels Institute of Science Technology & Advanced Studies (VISTAS), Chennai, Tamilnadu, India. (e-mail: abiramy.rv@gmail.com)

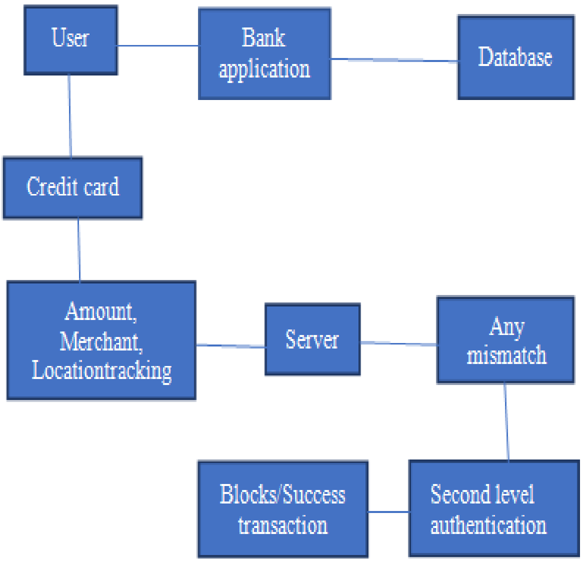
Kumar Narayanan, Department of Computer Science & Engineering, Vels Institute of Science Technology & Advanced Studies (VISTAS), Chennai, Tamilnadu, India.

R. Anandan, Department of Computer Science & Engineering, Vels Institute of Science Technology & Advanced Studies (VISTAS), Chennai, Tamilnadu, India.

C. Swaraj Paul, Department of Computer Science & Engineering, Vels Institute of Science Technology & Advanced Studies (VISTAS), Chennai, Tamilnadu, India.

IV. LIST OF MODULES

1 Account Creation
2 Generate transaction Data.
3 Identify the Transaction
4 Block the Fraud transaction


User
Bank application
Database
Credit card
Amount, Merchant, Locationtracking
Server
Any mismatch
Blocks/Success transaction
Second level authentication

System architecture

1.Account Creation

Initial user needs to create his own account and register his credit card detail in the server. Once server registers the user then server will provide pin number to the user to access the credit card and load that information on database server. Database server will maintain the user personal detail and all the transaction detail that are processed by the user.

2.Generate transaction Data

Initial user needs to provide some transaction data. Once the server approves the transaction that information will be loaded to the database server. The transaction information includes transaction location, purchased amount, merchant details, etc. all these data will be loaded in the server based on the unique id of the customers card number (CCNO).

3.Identify the Transaction

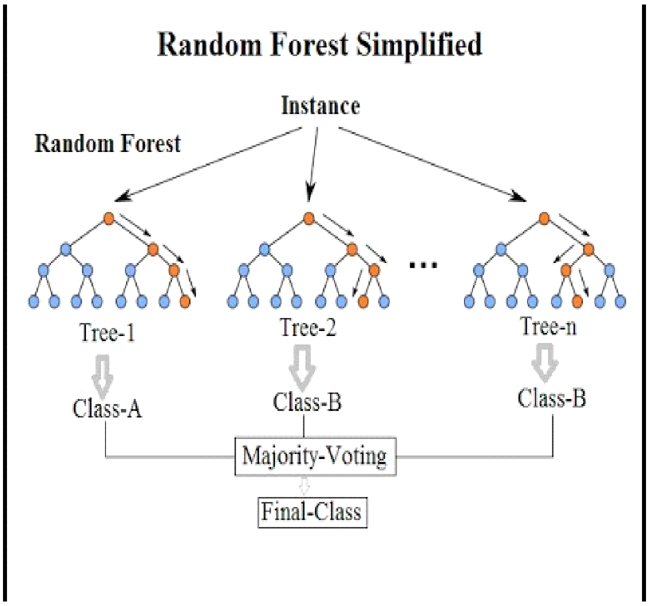
In Order to identify the user transaction, we introduce two categories of fraudulent transactions:

- Fraud detection due to impostors and
- Transactions occurred in past yet erratic with the new transaction location.

Server will cross verify the truncation, processed by the user, i.e., the server will verify whether the card holder had done any purchase in that particular location or not. Then cross verifies the past history with that merchant based on amount and product purchased which can be identified by Random forest algorithm.

Random Forest is an ensemble method. Random forest is a forest of multiple decision trees based on the individual predictions, whose observations are accurate than an individual-base classifier [1]. This method produces a single classifier by merging many diverse independent-base classifiers which is also called bagging. This technique lowers the risk of overfitting.

1. Randomly select “k” features from total “m” features. (Where k << m**)**
2. Using the best split approach find the root node among the “k” features.
3. Again, implement the best split to determine the daughter nodes “d”.
4. Repeat step1, step2 and step3 until “1” number of nodes is achieved.
5. Construct a forest by repeating the steps 1 to step 4 several times until we form “n” number of trees.



(Google source)

4.Block the Fraud transaction

For each transaction the server will cross verify the user’s transaction history. Then by using the feedback mechanism, it sends an OTP to the user registered mobile number or by asking conformation from the user. As a second level of authentication if the user confirms his identity then the transaction will process otherwise server will block that particular transaction. Thus, the transaction is identified as fraud and cardholders’ profile is updated.

V. EXPERIMENT

In decision tree the nodes represent the data rather than the decision. Hence also called classification tree [12]. Each branch in the tree consist of a set of rules named as decision rules declared with an if-else clause.
To achieve high accuracy multiple trees are combined together in ensemble method:

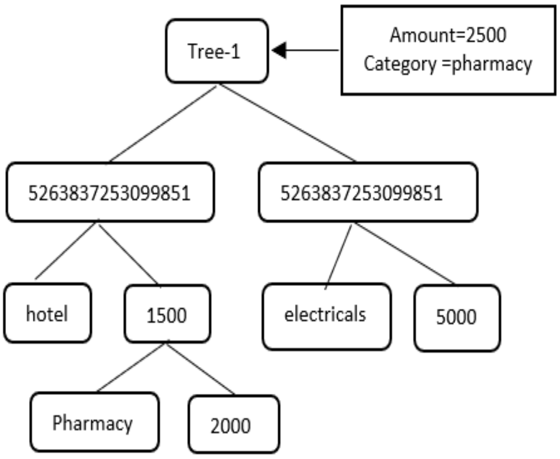
- ➢ A random forest classifier is a forest of many trees intended to rise the classification rate.
- ➢ Bagging constructs multiple trees by resampling the source data to achieve consensus.

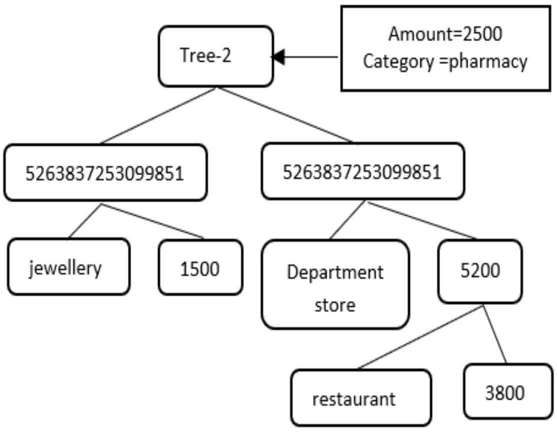
A random forest is build using the transaction amount, merchant category and transaction location. The decision trees now cross check the classified data with the respective cardholder’s database [7]. If the data’s mismatch a second level authentication is performed through feedback mechanism.

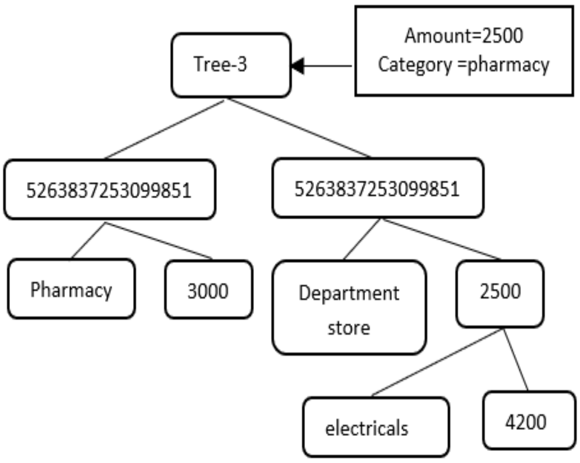
Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

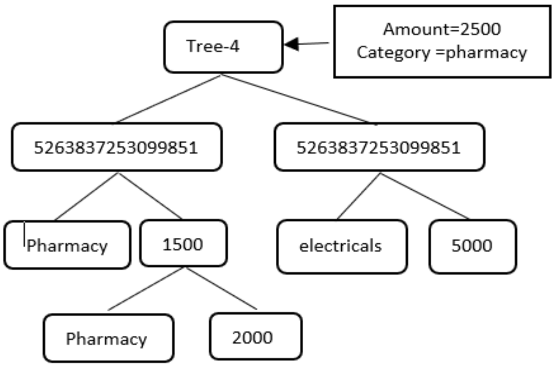
1.purchase amount

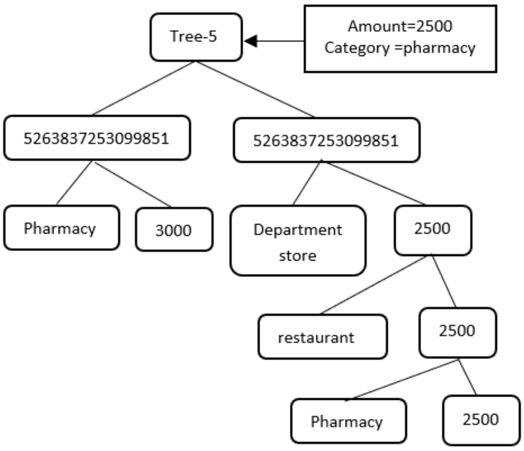
Variables – merchant category and price.

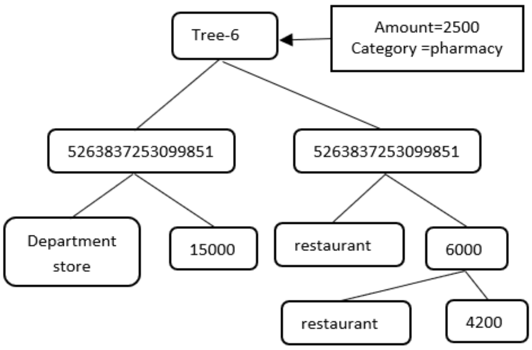


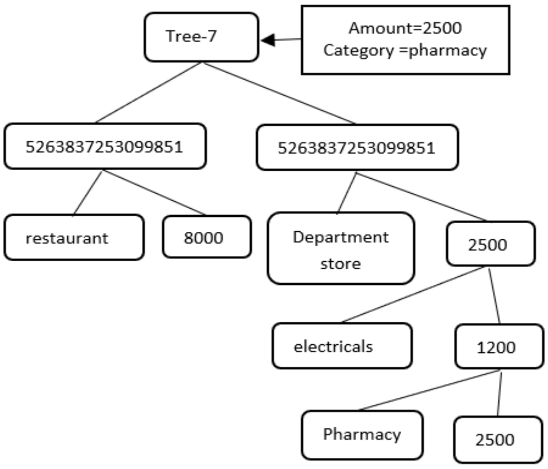










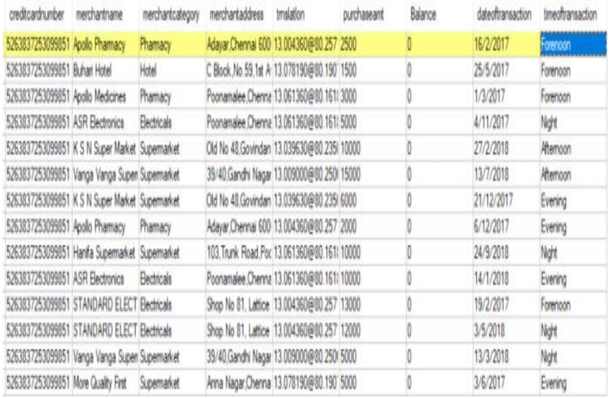


Bagging from all the above trees gives, pharmacy with [2000,2500,3000] maximum times than other amounts thus when number of decision trees get increased the prediction will be more accurate. Here from these inputs the amount 2500 belongs to the subset {2000,2500,3000} so the new transaction amount will be allowed by our random forest algorithm. Advantages of decision trees in machine learning:

- Accurate results regardless of violating the source data assumptions.
- The usage of additional data points declines the cost of predicting data.
- White box model makes outcomes simple to illustrate and understand.
- A tree’s accuracy can be examined and assessed.
- Works for both unqualified or mathematical data.
- Can design puzzles with multiple results.

VI. EVALUATION

Input 1: creditcardnumber = 5263837253099851
merchant category = pharmacy
merchant name = Apollo Pharmacy
latitudelongitude = 13.004360@80.257130
amount = 2500
time = forenoon



creditcardnumber	merchantname	merchantcategory	merchantaddress	lmlation	purchaseamt	Balance	dateoftransaction	timeoftransaction
5263837253099851	Apollo Phamacy	Phamacy	Adayar,Chennai 600	13.004360@80.257	2500	0	16/2/2017	Forenoon
5263837253099851	Buhari Hotel	Hotel	C Block,No.59,1st A	13.078190@80.190	1500	0	25/5/2017	Forenoon
5263837253099851	Apollo Medicines	Phamacy	Poonamalee,Chenna	13.061360@80.161	3000	0	1/3/2017	Forenoon
5263837253099851	ASR Electronics	Electricals	Poonamalee,Chenna	13.061360@80.161	5000	0	4/11/2017	Night
5263837253099851	K S N Super Market	Supemarket	Old No 48,Govindan	13.039630@80.235	10000	0	27/2/2018	Afternoon
5263837253099851	Vanga Vanga Super	Supemarket	39/40,Gandhi Nagar	13.009000@80.250	15000	0	13/7/2018	Afternoon
5263837253099851	K S N Super Market	Supemarket	Old No 48,Govindan	13.039630@80.235	6000	0	21/12/2017	Evening
5263837253099851	Apollo Phamacy	Phamacy	Adayar,Chennai 600	13.004360@80.257	2000	0	6/12/2017	Evening
5263837253099851	Hanifa Supemarket	Supemarket	103,Trunk Road,Poo	13.061360@80.161	10000	0	24/9/2018	Night
5263837253099851	ASR Electronics	Electricals	Poonamalee,Chenna	13.061360@80.161	10000	0	14/1/2018	Evening
5263837253099851	STANDARD ELECT	Electricals	Shop No 81, Lattice	13.004360@80.257	13000	0	19/2/2017	Forenoon
5263837253099851	STANDARD ELECT	Electricals	Shop No 81, Lattice	13.004360@80.257	12000	0	3/5/2018	Night
5263837253099851	Vanga Vanga Super	Supemarket	39/40,Gandhi Nagar	13.009000@80.250	5000	0	13/3/2018	Night
5263837253099851	More Quality First	Supemarket	Anna Nagar,Chenna	13.078190@80.190	5000	0	3/6/2017	Evening

1.Purchace amount comparison

Total amount in previous transaction belongs to the category“pharmacy” [2500,3000,2000]comparing current amount with each amount in pharmacy.

```
1.) 2500<=2500||2500>=2500
2.)2500<=3000||2500>=3000
3.)2500<=2000||2500>=2000
{
…//count of maximum
Treeset<Integer> set =new Treeset();
return count;
}
```

The count is '2' and from 3 values two amount values present inside the bound. There the decision tree for amount in pharmacy category will range from (2000<= X >=3000) where X=2500;

2.Location comparison

String latlng = 13.004360@80.257130
{
Using Select * query we retrieve all location values corresponding to creditcard number '5263837253099851'

```
...//latitude longitude bounds
String[] locarray =latlng.split("@");
 Treeset<Double> latset =new Treeset();
 Treeset<Double> longset =new Treeset();
latset.add(Integer.parseInt(locarray[0]));
longset.add(Integer.parseInt(locarray[1]));
latset.add(...);
longset.add(...);
}
```

Using the Collection command in Treeset we can retrieve the upper bound value.

```
latset.getLast();
latset.getFirst();
```

These two commands retrieve the least minimum value and the top maximum value.

...// Iterate the Collection and get the count of maximum values getting nearby to the maximum and the minimum value.

If the count value is high comparing to the average count of location, then the current location will be considered to the transaction limits.When all the parameters satisfy the condition, the current transaction will be allowed, and the transaction details will get updated.

Input 2: creditcardnum = 5263837253099851
merchant category = pharmacy
merchant name = Apollo Pharmacy
latitude longitude = 13.004360@80.257130
amount = 3500
time = forenoon

1.Purchace amount comparison

Total amounts in previous transaction belongs to the category " pharmacy" [2500,3000,2000]comparing current amount with each amount in pharmacy.

```
1.) 3500<=2500||3500>=2500
2.)3500<=3000||3500>=3000
3.)3500<=2000||3500>=2000
{
....//count of maximum
Treeset<Integer> set =new Treeset();
return count;
}
```

The count is '2' and from 3 values two amount values present inside the bound. There the decision tree for amount in pharmacy category will range from (X >=3000) where X=3500;

2.Location comparison

String latlng = 13.004360@80.257130
{
Using Select * query we retrieve all location values corresponding to creditcard number '5263837253099851'

```
...//latitude longitude bounds
String[] locarray =latlng.split("@");
 Treeset<Double> latset =new Treeset();
 Treeset<Double> longset =new Treeset();
latset.add(Integer.parseInt(locarray[0]));
```

longset.add(Integer.parseInt(locarray[1]));
latset.add(...);
longset.add(...);
}

Using the Collection command in Treeset we can retrieve the upper bound value.

latset.getLast();
latset.getFirst();

These two commands retrieve the least minimum value and the top maximum value.

...// Iterate the Collection and get the count of maximum values getting nearby to the maximum and the minimum value.

Here the amount is varied for a certain limit. And all the other parameters satisfy the upper and lower limits of the values present in the previous transaction dataset.

Due to the variation in amount the current transaction will be allowed for first level authentication.

- ➢ Pin Number Verification

When the pin number verification get success then transaction details will get updated.

Input 3: creditcardnum = 5263837253099851
merchant category = pharmacy
merchant name = Apollo Pharmacy
latitude longitude = 16.004360@83.255130
amount = 25000
time = forenoon

1.Purchace amount comparison

Total amounts in previous transaction belongs to the category " pharmacy” [3500,2500,3000,2000]comparing current amount with each amount in pharmacy.

1.) 25000<=2500||25000>=2500
2.)25000<=3000||25000>=3000
3.)25000<=2000||25000>=2000
4.)25000<=3500||25000>=3500
{
....//count of maximum
Treeset<Integer> set =new Treeset();
return count;
}

The count is '4' and from all the values '25000' exceeding the limit of lower bound '2500' and upper bound '3500'

2.Location comparison

String latlng = 16.004360@83.255130
{

Using Select * query we retrieve all location values correspondingto creditcard number '5263837253099851'

...//latitude longitude bounds
String[] locarray =latlng.split("@");
Treeset<Double> latset =new Treeset();
Treeset<Double> longset =new Treeset();
latset.add(Integer.parseInt(locarray[0]));
longset.add(Integer.parseInt(locarray[1]));
latset.add(...);
longset.add(...);
}

Using the Collection command in Treeset we can retrieve the upper bound using

latset.getLast();
latset.getFirst();

These two commands retrieve the least minimum value and the top maximum value. Iterate the Collection and get the count of maximum values getting nearby to the maximum and the minimum value. If the count value is high compared to the average count of location, then the current location will be considered as the transaction limits. Here the latitude value is exceeding the upper bound limit, so our algorithm logic triggers the second level authentication.

1. Pin number Verification
2.Custom OTP Verification

When these two authentications get success, transaction will be done, and the transaction details will get updated.

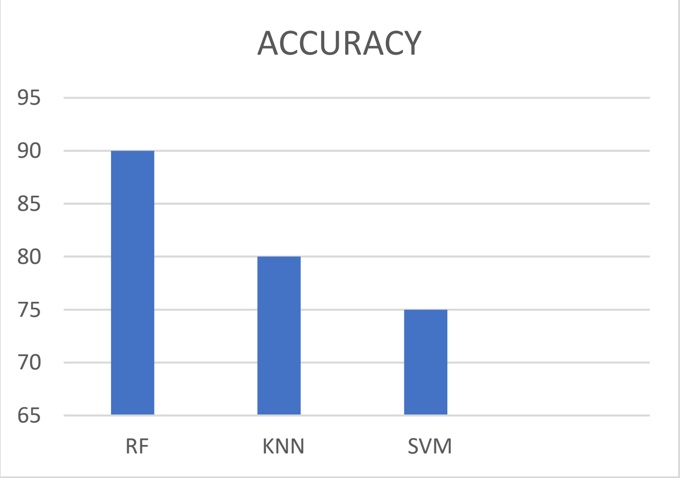
VII. ADVANTAGES

- Data mining helps bank to maintain their credit card customers.
- Credit card userscan overcome the deficits because of fake events.
- Reduction in number of fraud transaction.
- Enhanced security to meet today’s threat.

VIII. RESULT AND ANALYSIS

The proposed method becomes difficult to test using real dataset. Since Banks do not concur to part their statics with experimenter. We can make use of the standard data file available on WWW analysis. We have tried all the cases related to credit card transaction. This suggest that the fraudsters are well versed with the card bearer’s performance whereas the impostors can get the past report of credit card [1]. Consequently, impostors proceeding cost is close to usual transactions, but the only disparity between them is the repetition of transaction location in a duration of time because at all the times the impostors want to profit at the earliest opportunity.

In the existing paper many classification algorithms (KNN,SVM,LR) have been used, which makes the project more complicated and difficult to understand. But random forest algorithm makes it easier by using bagging concept with accurate prediction. Feedback mechanism together with RF provides faster as well as high security against frauds.



Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

IX. CONCLUSION

Clearly, credit card fraud is an act of culprit falsehood. Hence this paper is very helpful to the recent frauds in credit card field. This article can overcome the several distinct kinds of fraud, for example bankruptcy fraud, counterfeit fraud, theft fraud, application fraud and behavioral fraud, and as studied estimates to identify them. On the basis of moral aspect, banks and credit card companies must undertake steps to spot all fraudulent cases. Else the bank shall face with moral difficulty. Whether to detect fraudulent cases, or pay attention to shareholder interests and avoid uneconomic costs? As a next phase of this research, we will target on the implementation of 'suspicious 'scorecard on real data-set and its evaluation. The idea is to extend the research on any Asian countries, probably India.

AKNOWLEDGEMENT

I would like to convey my utmost credits to my guide Dr.N. Kumar Narayanan for his patient guidance throughout the study work. And I would also like to thank Dr. R. Anandan and Mr. C. Swaraj Paul for their constant supervision and by providing necessary information for my research.

REFERENCES

1. Changjun Jiang, Jiahui Song, Guanjun Liu, Member, IEEE, Lutao Zheng, and Wenjing Luan "Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism" IEEE Internet of Things Journal, vol., no., March 2018.
2. Lutao Zheng, Guanjun Liu, Chungang Yan, Changjun Jiang. "Transaction Fraud Detection Based on Total Order Relation and Behavior Diversity", IEEE Transactions on Computational Social Systems, 2018.
3. M. U. Sapozhnikova, A. V. Nikonov, A. M. Vulfin, M. M. Gayanova, K. V. Mironov, D. V. Kurennov. "Anti-fraud system on the basis of data mining technologies", 2017.
4. Alejandro Correa Bahnsen, Djamila Aouada, Aleksandar Stojanovic, Bjorn Ottersten. "Feature engineering strategies for credit card fraud detection", Expert Systems with Applications, 2016.
5. Bolton, R. J. and Hand, D. J. 2002. Statistical fraud detection: A review. Statistical Science 28(3):235-255.
6. David J. Wetson, David J.Hand,M Adams,Whitrow and Piotr Jusczak "Plastic Card Fraud Detection using Peer Group Analysis" Springer, Issue 2008.
7. Aleandro correa bahnsen, AleksandarStojanovic, DjamilaAouada and BjornOttersten" Cost Sensitive Credit Card Fraud Detection using Bayes Minimum Risk".
8. Maguire S. 2002. Identifying Risks During Information System Development: Managing the Process, Information Management & Computer Security, 10(3): 126–134.
9. Duncan M D G. 1995. The Future Threat of Credit Card Crime, RCMP Gazette, 57 (10): 25–26.
10. Evandro Caldeira, Gabriel Brandao, Adriano C. M. Pereira. "Characterizing and preventing chargebacks in next generation web payments services", 2012 Fourth International Conference on Computational Aspects of Social Networks (CASoN), 2012.
11. Breiman. L. Friedman, Isbell. R. & Stone, C. (1984). "Classification and regression trees". Wadsworth International Group.
12. J.A. Hartigan, and M.A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," Journal of the Royal Statistical Society, vol. 28, no. 1, pp. 100-108, 1979.
13. Philippe Fournier-Viger blog[online] Posted on 2018-12-03 by P. Fournier-VigerAvailablehttp://www.data-mining.philippe-fournier viger.com/datamining Decision trees for classification [online]. Availablehttps://www.lucidchart.com/decision
14. Web service-based communication [online] Availablehttps://www.java4s.com/web-services

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)