International Journal of Research
Available at https://edupediapublications.org/journals

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 12
April 2018

ATM Card Fraud Detection System Using Machine Learning Techniques

Dr. B. Prakash#1, G. Venu Madhava Murthy#2, P. Ashok#3, B. Pavan Prithvi#4, S. Sai Harsha Kira#5
[1]*professor, HOD, #1-5Department of CSE, #1-5Vignan Institute of Technology and Science*

[1]prakash.vza@gmail.com , [2]venumadhav.gvmm@gmail.com ,[3]ashok.poolla6@gmail.com

[4]pavanboddu17@gmail.com , [5]harsha1097@gmail.com

Abstract— Financial fraud is an ever-growing menace with far consequences in the financial industry. ATM card fraud detection, which is a data mining problem, becomes challenging due to two major reasons. First, the profiles of normal and fraudulent behaviour change constantly and secondly, ATM card fraud data sets are highly skewed. The performance of fraud detection in ATM card transactions is greatly affected by the sampling approach on dataset, selection of variables and detection technique used. ATM card fraud causes disruptions in the digital payment and banking sector. Machine Learning is quickly emerging as the standard for mitigating risks occurring due to the use of ATM cards. This paper explores the performance of Decision Tree, Logistic Regression on largely imbalanced data sourced from European cardholders containing over 2,00,000 transactions. The work is implemented in R language using RStudio along with a GUI application developed using Shiny, which is a framework for writing rich web apps using R.

Keywords— **ATM Fraud, Decision Tree, Shiny, Detection, Transaction, Machine Learning**

I. INTRODUCTION

In recent years, the prevailing data mining concerns people with ATM card fraud detection model based on data mining. [3] Data mining had played an imperative role in the detection of ATM card fraud in online transactions. Since our problem is approached as a classification problem, classical data mining algorithms are not directly applicable. [5][7] So, an alternative approach is made by using general purpose meta heuristic approaches like machine learning techniques. This project is to propose an ATM card fraud detection system using genetic algorithm. Machine learning techniques are evolutionary algorithms which aim at obtaining better solutions as time progresses. [6] It aims in minimizing the false alerts using machine learning techniques where a set of interval valued parameters are optimized. To develop an ATM card fraud detection system using genetic algorithm. During the ATM card transaction, the fraud is detected, and the number of false alert is being minimized by using genetic algorithm. [4] Instead of maximizing the numbers of correctly classified transactions we defined an objective function where the misclassification costs are variable and thus, correct classification of some transactions are more important than correctly classifying the others. This information regarding analysis done for the proposed system. [1][2] Here the goal of the project is explained, and the cost and performance factors which will affect the feasibility of the project is explained. Fraud detection based on the analysis of existing purchase data of cardholder is promising way to reduce the rate of successful ATM card frauds. [8] Since humans tend to exhibit specific behaviourist profiles, every cardholder can be represented by a set of patterns containing information about the typical purchase category, the time since the last purchase, the amount of money spent, etc. Deviation from such patterns is a potential threat to the system. [9] Fraudulent transactions are a problem encompassing digital payment system, thereby suggesting a need to develop and dissolve novel solutions for this case RStudio is a statistical and graphical programming language. RStudio has deeply integrated R programming language and allows to develop programs for Data Science field estimation.

II. MACHINE LEARNING

Machine learning algorithms are often categorized as supervised, unsupervised and semi supervised. Supervised machine learning algorithm: can apply what has been learned in the past to new data using labelled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system can provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors to modify the model accordingly. Unsupervised machine learning algorithm: are used when the information used to train is neither classified nor labelled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabelled data. The system doesn’t figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabelled data, Semi-supervised machine learning algorithm: fall somewhere in between supervised and unsupervised learning, since they use both labelled and unlabelled data for training – typically a small amount of labelled data and a large amount of unlabelled data. The systems that use this method can considerably improve learning accuracy. Usually, semi-supervised learning is chosen when the acquired labelled data requires skilled and relevant resources to train it / learn from it. Otherwise, acquiring unlabelled data generally doesn’t require additional resources

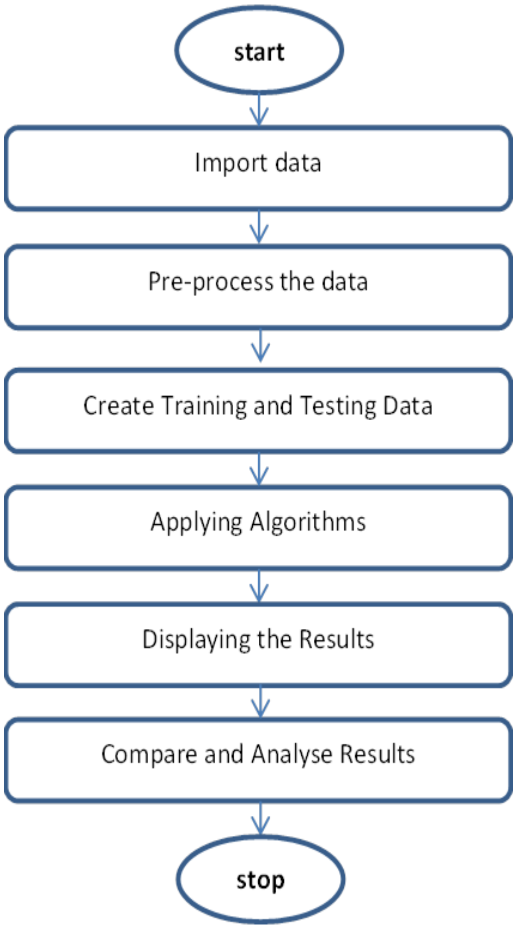


Fig 1. Block diagram of the current system

III. IMPLEMENTATION

Dataset Description: The dataset contains transactions made by ATM cards of the customers of a foreign bank. This dataset presents transactions that occurred in 48 hours, where we have 492 frauds out of 2,84,807 transactions. The dataset is highly unbalanced and the positive class (frauds) account for 0.172% of all transactions. It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant and cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise,

A. Algorithms:

Step wise Procedure for Decision tree algorithm

Step 1: Imported the data

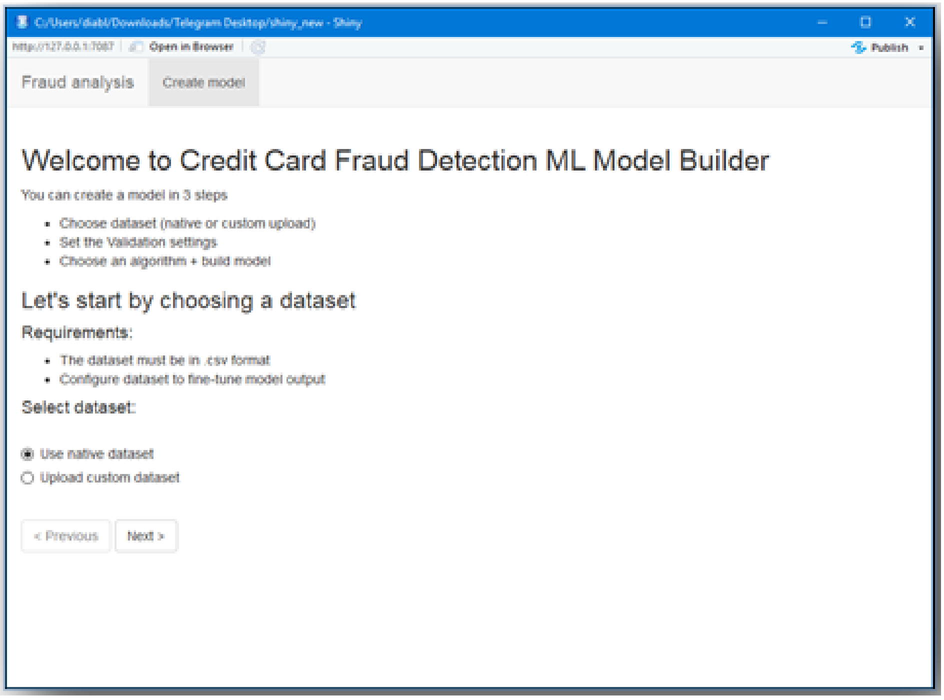
Step 2: Set the ratio for training and testing data

Step 3: Split dataset based on the most significant attribute chosen to be the root

Step 4: Associated number of buckets for build parameters

Step 5: Generated confusion matrix and underlying accuracy

Decision Tree: Decision tree is a graph to represent choices and their results in form of a tree. The nodes in the graph represent an event or choice and the edges of the graph represent the decision rules or conditions. It is mostly used in Machine Learning and Data Mining applications using R. This stands the best algorithm for this problem after machine learning analysis. Graph may be displayed in two ways based on the values. We may get graph in horizontal or in a vertical way. Sometimes graph may get very big graph. In such case we may implement additional features in order to reduce the size of the graph.



Welcome to Credit Card Fraud Detection ML Model Builder

You can create a model in 3 steps

- Choose dataset (native or custom upload)
- Set the Validation settings
- Choose an algorithm + build model

Let's start by choosing a dataset

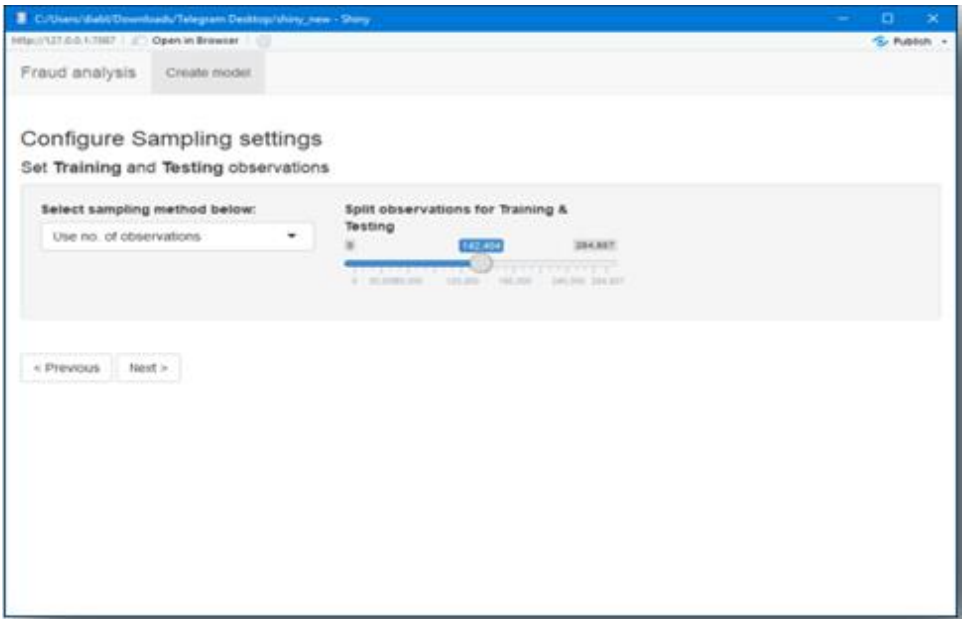
Requirements:

- The dataset must be in .csv format
- Configure dataset to fine-tune model output

Select dataset:

● Use native dataset
○ Upload custom dataset

Fig 2. GUI of the current system



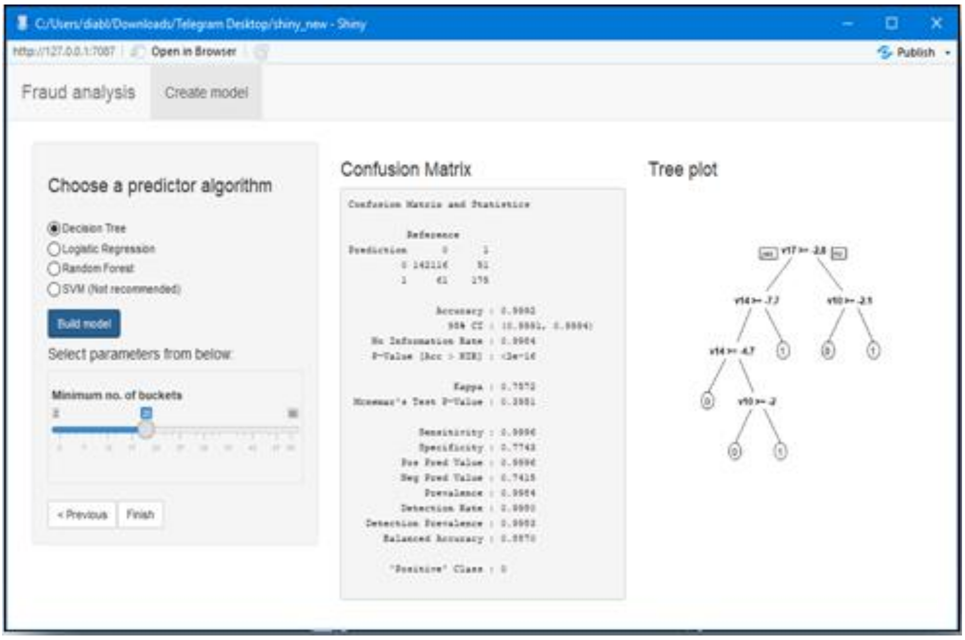
Configure Sampling settings

Set **Training** and **Testing** observations

Select sampling method below: Use no. of observations

Split observations for Training & Testing: 142,404

Fig.3 GUI of Training and Testing Data



Choose a predictor algorithm

- ● Decision Tree
- ○ Logistic Regression
- ○ Random Forest
- ○ SVM (Not recommended)

Select parameters from below

Minimum no. of buckets

Confusion Matrix

```
Confusion Matrix and Statistics

          Reference
Prediction      0      1
         0 142116     51
         1     61    175

               Accuracy : 0.9992
                 95% CI : (0.9991, 0.9994)
    No Information Rate : 0.9984
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7572
 Mcnemar's Test P-Value : 0.2881

            Sensitivity : 0.9996
            Specificity : 0.7743
         Pos Pred Value : 0.9996
         Neg Pred Value : 0.7415
             Prevalence : 0.9984
         Detection Rate : 0.9980
   Detection Prevalence : 0.9983
      Balanced Accuracy : 0.8870

       'Positive' Class : 0
```

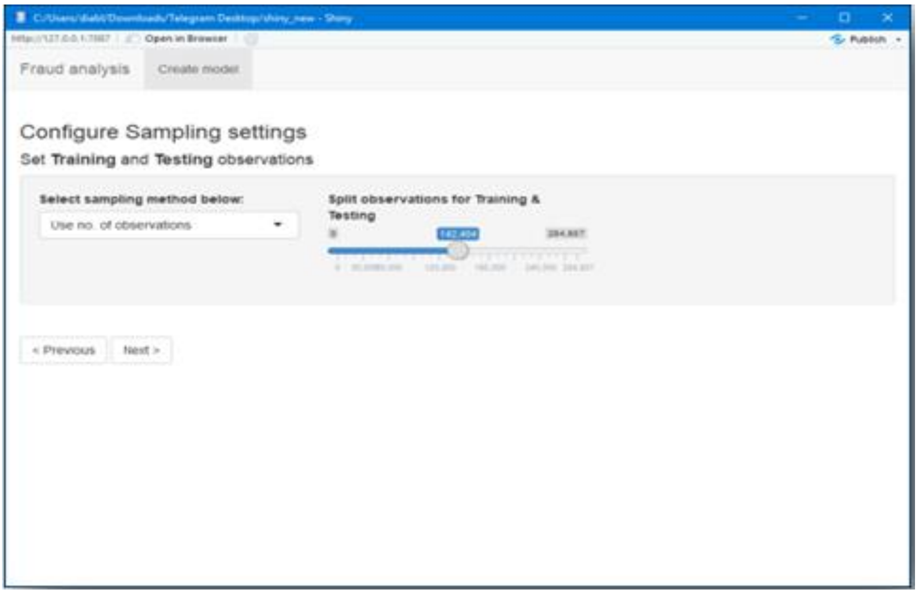
Fig 4. Confusion Matrix and Decision Tree

The above figures represent the output screens. This screen specifically shows the output of decision tree algorithm applied on the native dataset. The consolidated list of results includes the following columns:

- Confusion Matrix
- Accuracy (Decision Tree: 99.92%)
- Kappa value
- Active Class
- Balanced Accuracy Value

Logistic Regression: Logistic regression is one of the most popular machine learning algorithms for binary classification. This is because it is a simple algorithm that performs very well on a wide range of problems

Algorithm:
Step 1: Imported the data and checked for class bias
Step 2: Created training and testing samples.
Step 3: Computed information value for important variables
Step 4: Built logit models and predicted on testing data
Step 5: Performed model diagnostics



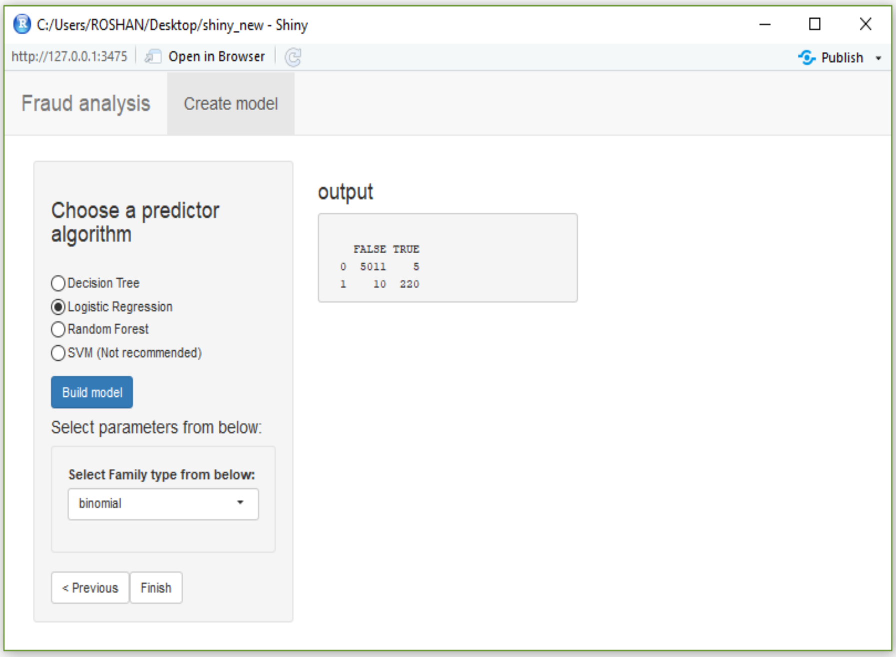
Configure Sampling settings

Set Training and Testing observations

Select sampling method below: Use no. of observations

Split observations for Training & Testing: 0 — 142,404 — 284,807

Fig 5. GUI of Train and Test Datasets



Choose a predictor algorithm

○ Decision Tree
● Logistic Regression
○ Random Forest
○ SVM (Not recommended)

Select parameters from below:

Select Family type from below: binomial

output

```
   FALSE TRUE
 0  5011    5
 1    10  220
```

Fig 6. Confusion Matrix of Logistic Regression

Comparative results:

The following table depicts the comparative performance of Logistic Regression matched against different algorithms. From the existing results, Support Vector Machine Technique gave an accuracy of 0.9520. By using Logistic Regression, Results seemed to be little better compared to SVM algorithm. By using Decision tree algorithm, Results obtained is 0.9992 and this seemed to be more promising when compared to the other two. The following table depicts the comparative results of the three algorithms

S.No	Algorithm	Accuracy
1	Decision Tree	0.9992
2	Logistic Regression	0.9790
3	Support Vector Machine	0.9520

Table1 : Comparative results

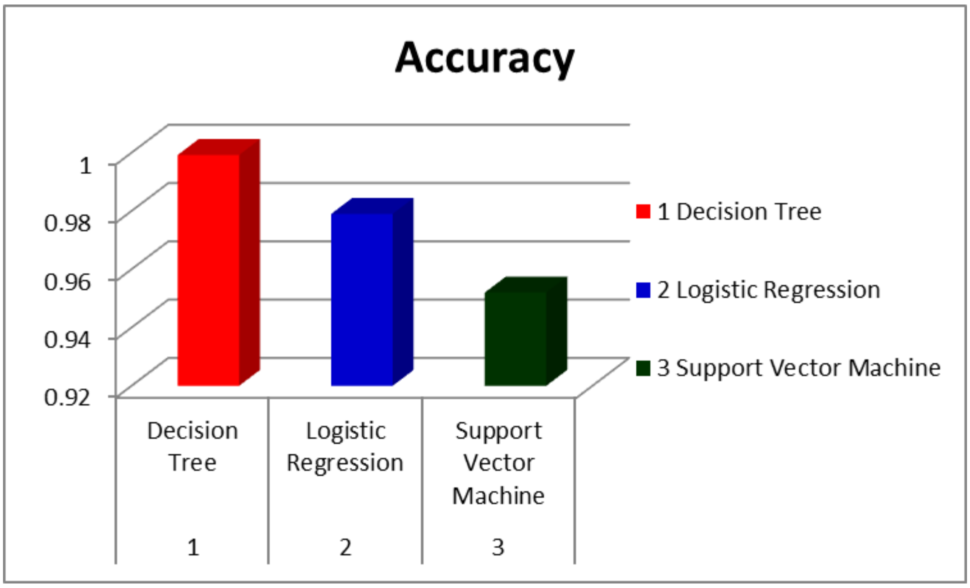


Fig 7. Graphical Representation of Result

IV. CONCLUSION

This method proves accurate in deducting fraudulent transaction and minimizing the number of false alert. Machine Learning techniques is a novel one in this literature in terms of application domain. If this algorithm is applied into bank ATM card fraud detection system, the probability of fraud transactions can be predicted soon after ATM card transactions. And a series of anti-fraud strategies can be adopted to prevent banks from great losses and reduce risks. The objective of the study was taken differently than the typical classification problems in that we had a variable misclassification cost. As the standard data mining algorithms does not fit well with this situation we decided to use multi population machine learning techniques to obtain an optimized parameter

V. FUTURE WORK

The findings obtained here may not be generalized to the global fraud detection problem. As future work, some effective algorithm which can perform well for the classification problem with variable misclassification costs could be developed. Future research includes but not limited to exploring more predictor algorithms and UX optimizations.

VII. ACKNOWLEDGEMENTS

Credits to RPubStudio for providing necessary dataset

VIII. REFERENCES

[1] Stolfo, S. J., Fan, D. W., Lee, W., Prodromidis, A., and Chan, P. K., 2000. CostBased Modeling for Fraud and Intrusion Detection: Results from the JAM Project, Proceedings of DARPA Information Survivability Conference and Exposition, vol. 2 (2000), pp. 130-144.

[2] Aleskerov, E., Freisleben, B., and Rao, B., 1997. CARDWATCH: A Neural Network Based Database Mining System for ATM Card Fraud Detection, Proceedings of IEEE/IAFE: Computational Intelligence for Financial Eng. (1997), pp. 220-226.

[3] M.J. Kim and T.S. Kim, “A Neural Classifier with Fraud Density Map for Effective ATM Card Fraud Detection,” Proc. Int’l Conf. Intelligent Data Eng. and Automated Learning, pp. 378-383, 2002.

[4]. Anderson M. (2008). _From Subprime Mortgages to Subprime ATM Cards‘. Communities and Banking, Federal Reserve Bank of Boston, pp. 21-23.

[5]. Anwer et al. (2009-2010). _Online ATM Card Fraud Prevention System for Developing Countries‘, International Journal of Reviews in Computing, ISSN: 2076-3328, Vol. 2, pp. 62- 70.

[6]. Arias, J.C. & Miller R. (2009). _Market Analysis of Student about ATM Cards‘. Business Intelligence Journal, Vol. 3, No. 1, pp. 23-36.

[7]. Bhatla T.P. et al. (2003). _Understanding ATM Card Frauds‘. Cards Business Review, 01, pp. 01-15.

[8]. Bhusari V. & Patil S. (2011). _Study of Hidden Markov Model in ATM Card Fraudulent Detection ‘. International Journal of Computer Applications, Vol. 20, No. 5, pp. 33-36.

[9]. Calem P. & Mester L. (1995). _Consumer Behavior and the Stickiness of ATM Card Interest Rates‘. The American Economic Review, Vol. 85, No. 5, pp. 1327-1333

[10]. Chakravorti S. (2003). _Theory of ATM Card Networks: A Survey of the Literature‘, Review of Network Economics, Vol. 2, Issue 2, pp. 50-68.