Enhanced System for Revealing Fraudulence in Credit Card Approval

B. Subashini [1], Dr. K. Chitra [2]

[1]Assistant Professor, Department of Computer Science,
V.V.Vanniaperumal College for Women, Virudhunagar.
[2]Assistant Professor, Department of Computer Science,
Government Arts College, Melur, Madurai.

Abstract

In developing countries like India, Bankers face more problems with the fraudsters. Data mining techniques are more useful to build a successful predictive model and visualize the report into meaningful information to the user. This research paper aims to enhance and evaluate the fraudulence in credit card approval process using the classification models based on decision trees (C5.0 & CART), Support Vector Machine (SVM) using SMO, BayesNet and Logistic Regression. Five methods to detect fraud are presented. Automatic credit card approval is the most significant process in the banking sector and financial institutions. This enhanced system prevents the fraud which is going to happen. So this paper proposes a good solution to the credit card approval using the above methods.

Keywords:- Credit card approval, Fraud, Data Mining, Classification, SVM, Logistic Regression

1. Introduction

Credit card fraud falls broadly into two categories: behavioral fraud and application fraud. Application fraud occurs when individuals obtain new credit cards from issuing companies using false personal information and then spend as much as possible in a short span of time [5]. In a move to curtail rising credit card frauds, the Reserve Bank of India has asked banks to bar international usage of debit and credit cards unless customers specifically ask for this feature. Banks have also been asked to enable blocking of cards through a text message request [6]. So nowadays, credit approval is the tremendous problem in the banking sector. Automatic credit approval is the process of granting credits or loans to customers. Prevention is better than cure. Fraud prevention is the proactive mechanism with the goal of disabling the occurrence of fraud.

This paper enhances the credit card approval process in the banking sector by using and comparing the classification methods such as decision trees, Support Vector Machine (SVM) and Logistic Regression. It depends on the performance metrics such as performance and accuracy.

The rest of the paper is organized as follows: Section 2 describes the classification methods which are used to apply in the credit card approval dataset and performance metrics. Section 3 presents experiment setup and results from various classification methods. Section 4 analyzes the performance of the classification methods' results. Section 5 concludes this work.

2. Data Mining

Data mining is the analysis step of knowledge discovery in databases. Data mining is a powerful new technology with great potential to help companies focus on the most important information in the data they have collected about the behavior of their customers and potential customers. Data mining derives its name from the similarities between searching for valuable information in a large database and mining a mountain for a vein of valuable ore. Specific uses of data mining include: Market segmentation, Customer churn, Fraud detection, Direct Marketing, Interactive marketing, Market basket analysis, Trend analysis [1] [2] [3] [4]. Three steps involved in the data mining process are Exploration, Pattern identification, Deployment. Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases [7]. In this paper, we enhance the credit card approval process to prevent fraud in the banking sector using classification methods.

2.1. Classification Methods

Classification is perhaps the most familiar and most popular data mining technique. Estimation and prediction may be viewed as types of classification. There are more classification methods such as statistical based, distance based, decision tree based, neural network based, rule based [8]. In this paper, we detect fraud using the classification algorithms C5.0, BayesNet, Classification Via Regression Trees (CART), Support Vector Machine using Sequential Minimal Optimization (SMO), Logistic Regression and we analyze their performance in fraud detection in the banking sector. The performance analysis is done on the basis of the following performance metrics:

a. Classified Instances - The importance performance measure is correctly classified instances and incorrectly classified instances.
b. ROC – ROC (Relative operating characteristic) curve shows the relationship between false and true positive.
c. Confusion Matrix – The confusion matrix illustrates the accuracy of the solution to a classification problem. The rows represent the actual classification and the

columns the predicted classification. In this matrix **Good classification denotes the Legitimate customer and Bad classification denotes the Fraud customer**.

3. Experiments and Results

For the experimental work in this paper, a dataset of credit card applications and approval decisions, Credit Card Approval, from UCI Repository of Machine Learning Databases and Domain Theories, was used. The dataset was used to detect fraudulent customer during their credit card approval process and to induce classification models for assessing credit card applications. The dataset has 20 Attributes (7 numerical, 13 categorical) plus the class label attribute. The dataset is interesting because there is a good mix of attributes: numerical, categorical with meaningful values. There are 1000 instances in this dataset.

The tests were made using the software Weka (Waikato Environment for Knowledge Analysis) which contains lot of classification algorithms. Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand [11].

I. C5.0

C5.0 builds decision trees from a set of training data in the same way as ID3, using the concept of Information entropy. The training data is a set $S=S_1,S_2,..$ of already classified samples. Each sample S_i consists of a p-dimensional vector $(x_{1,i},x_{2,i}, \ldots, x_{p,i})$, where the x_j represent attributes or features of the sample, as well as the class in which s_i falls. At each node of the tree, C5.0 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C5.0 algorithm then recurses on the smaller sublists. Gain is computed to estimate the gain produced by a split over an attribute. The gain of information is used to create small decision trees that can identify the answers with a few questions [9].

The following figure.1 illustrates the knowledge flow analysis using cross validation fold 10, and uses J48 algorithm for classification.

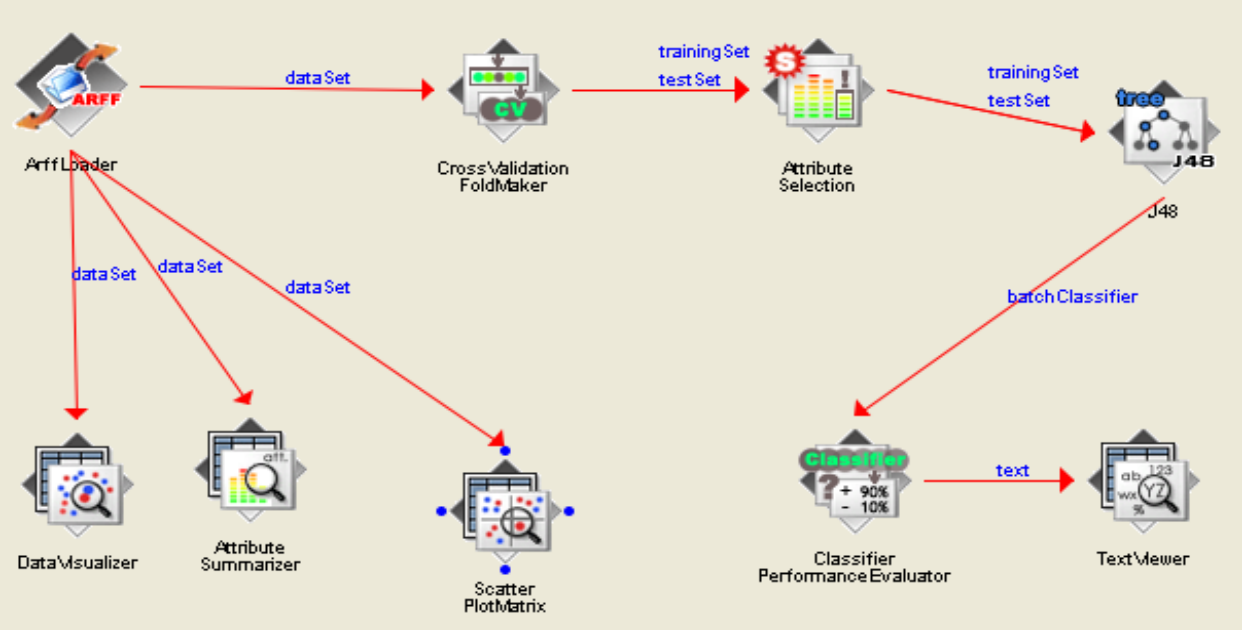

Figure 1. Knowledge Flow Analysis using J48 Algorithm

For the application in a decision tree, the algorithm used was J48, which is Java implementation of C5.0. Here, the training algorithm took only 0.02 seconds to classify 1000 instances. The classifier output for C5.0 is follows.

=== Evaluation result ===

Scheme: J48

Options: -C 0.25 -M 2

Correctly Classified Instances	724	72.4 %
Incorrectly Classified Instances	276	27.6 %
Kappa statistic	0.2988	
Mean absolute error	0.3389	
Root mean squared error	0.4306	
Relative absolute error	80.5279 %	
Root relative squared error	93.8224 %	
Coverage of cases (0.95 level)	99.1 %	
Mean rel. region size (0.95 level)	98.2 %	
Total Number of Instances	1000	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.850	0.570	0.777	0.850	0.812	0.303	0.700	0.794	good
	0.430	0.150	0.551	0.430	0.483	0.303	0.700	0.536	bad
Weighted Avg.	0.724	0.444	0.709	0.724	0.713	0.303	0.700	0.717	

=== Confusion Matrix ===

a	b	<-- classified as
595	105	a = good
171	129	b = bad

II. CART

A CART tree is a binary decision tree that is constructed by splitting a node into two child nodes repeatedly, beginning with the root node that contains the whole learning sample. Used by the CART (classification and regression tree) algorithm, Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset. Gini impurity can be computed by summing the probability of each item being chosen times the probability of a mistake in categorizing that item. It reaches its minimum (zero) when all cases in the node fall into a single target category.

=== Evaluation result ===

Scheme: ClassificationViaRegression

Correctly Classified Instances	741	74.1 %
Incorrectly Classified Instances	259	25.9 %
Kappa statistic	0.297	
Mean absolute error	0.3447	
Root mean squared error	0.4186	
Relative absolute error	81.8997 %	
Root relative squared error	91.2062 %	
Coverage of cases (0.95 level)	99.1 %	
Mean rel. region size (0.95 level)	94.35 %	

Total Number of Instances 1000

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.907	0.647	0.766	0.907	0.831	0.317	0.752	0.866	good
0.353	0.093	0.620	0.353	0.450	0.317	0.752	0.563	bad
Weighted Avg. 0.741	0.481	0.722	0.741	0.716	0.317	0.752	0.775	

=== Confusion Matrix ===

a	b	<-- classified as
635	65	a = good
194	106	b = bad

III. Support Vector Machine(SVM)

In machine learning, the polynomial kernel is a kernel function commonly used with support vector machines (SVMs) and other kernelized models, that represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables. For degree-*d* polynomials, the polynomial kernel is defined as $K(x,y)= (x^T y+c)^d$ where x and y are vectors in the input space, i.e. vectors of features computed from training or test samples, c >0 is a constant trading off the influence of higher-order versus lower-order terms in the polynomial. In this paper SVM is implemented using Sequential Minimal Optimization (SMO).

=== Evaluation result ===

Scheme: SMO

Correctly Classified Instances	724	72.4 %
Incorrectly Classified Instances	276	27.6 %
Kappa statistic	0.2032	
Mean absolute error	0.276	
Root mean squared error	0.5254	
Relative absolute error	65.5855 %	

Root relative squared error 114.4562 %

Coverage of cases (0.95 level) 72.4 %

Mean rel. region size (0.95 level) 50 %

Total Number of Instances 1000

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.934	0.767	0.740	0.934	0.826	0.240	0.584	0.737	good
0.233	0.066	0.603	0.233	0.337	0.240	0.584	0.371	bad
Weighted Avg. 0.724	0.556	0.699	0.724	0.679	0.240	0.584	0.627	

=== Confusion Matrix ===

a	b	<-- classified as
654	46	a = good
230	70	b = bad

IV. Logistic Regression

Logistic regression or logit regression is a type of regression analysis used for predicting the outcome of a categorical dependent variable based on one or more predictor variables. Instead of fitting the data to a straight line, logistic regression uses a logistic curve. The formula for a univariate logistic curve is

$$p = \frac{e^{c_0+c_1x_1}}{1+e^{c_0+c_1x_1}}$$

To perform the logarithmic function can be applied to obtain the logistic function

$$log_e = \frac{p}{1-p} = c_0+c_1x_{10}$$

Logistic regression is simple, easy to implement, and provide good performance on a wide variety of problems [10].

=== Evaluation result ===

Scheme: Logistic

Correctly Classified Instances 731 73.1 %

Incorrectly Classified Instances 269 26.9 %

Kappa statistic 0.2807

Mean absolute error 0.3406

Root mean squared error 0.4178

Relative absolute error 80.9379 %

Root relative squared error 91.0334 %

Coverage of cases (0.95 level) 99.9 %

Mean rel. region size (0.95 level) 97.55 %

Total Number of Instances 1000

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.890	0.640	0.764	0.890	0.822	0.295	0.755	0.870	good
0.360	0.110	0.584	0.360	0.445	0.295	0.755	0.564	bad
Weighted Avg.	0.731	0.481	0.710	0.731	0.709	0.295	0.755	0.778

=== Confusion Matrix ===

a	b	<-- classified as
623	77	a = good
192	108	b = bad

V. BayesNet

Bayes Nets or Bayesian networks are graphical representation for probabilistic relationships among a set of random variables. Given a finite set $X = \{ X_1, \ldots, X_n\}$ of discrete random variables where each variable X_i may take values from a finite set, denoted by $Val(X_i)$. A Bayesian network is an annotated directed acyclic graph (DAG) G that encodes a joint probability distribution over X. The nodes of the graph correspond to the random variables $X_1, \ldots, X_n$. The links of the graph correspond to the direct influence from one

variable to the other. If there is a directed link from variable X_i to variable X_j, variable X_i will be a parent of variable X_j. Each node is annotated with a conditional probability distribution (CPD) that represents $p(X_i \mid P_a(X_i))$, where $P_a(X_i)$ denotes the parents of X_i in G. The pair (G, CPD) encodes the joint distribution $p(X_1, \ldots, X_n)$. A unique joint probability distribution over X from G is factorized as:

$$p(X_1,\ldots,X_n) = \prod_i (p(X_i \mid Pa(X_i)))$$

=== Evaluation result ===

Scheme: BayesNet

Correctly Classified Instances	724	72.4 %
Incorrectly Classified Instances	276	27.6 %
Kappa statistic	0.2857	
Mean absolute error	0.3386	
Root mean squared error	0.424	
Relative absolute error	80.4581 %	
Root relative squared error	92.3654 %	
Coverage of cases (0.95 level)	99.4 %	
Mean rel. region size (0.95 level)	94.75 %	
Total Number of Instances	1000	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.863	0.600	0.770	0.863	0.814	0.293	0.748	0.868	good
0.400	0.137	0.556	0.400	0.465	0.293	0.748	0.535	bad
Weighted Avg. 0.724	0.461	0.706	0.724	0.709	0.293	0.748	0.768	

=== Confusion Matrix ===

a b <-- classified as

604 96 | a = good

180 120 | b = bad

4. Performance Analysis

The chosen methods to build classifier models are C5.0, CART from decision tree methods, SVMs using SMO algorithm with kernels of polynomial functions, Logistic Regression and BayesNet. All these methods are used to detect fraud in the banking sector using Credit Card Fraud data set. The Table 1 summarizes the classification values and success rate for all five classification methods C5.0, BayesNet, Classification via Regression Trees (CART), Support Vector Machine using Sequential Minimal Optimization (SMO), Logistic Regression and we analyze their performance in fraud detection in the banking sector. Here, Good denotes the Legitimate User and Bad denotes the Fraud User.

C5.0 using J48, SVM using SMO, BayesNet are giving the success rate of 72.4% whereas the Bad → Good classification is more in SVM using SMO. Because classifying a Bad customer as Good is worse than classifying a Good customer as Bad. Logistic Regression method is providing 73.1% success rate and CART gets the highest success rate 74.1%. Hence it is shown here that depending upon the success rate CART outperforms the other models whereas considering the Bad → Good classification J48 shows better performance. Any financial institution needs to retain its customers. A legitimate customer must not be classified as Fraud or a Fraud must not be classified as Legitimate. Hence It is understood from this work that while making decisions on classifying customers combination of different classification models need to be used to make correct decision about a customer.

Table 1. Performance Analysis of Classification Methods

	J48	*Logistic*	*SMO*	*BayesNet*	*CART*
Good-->Good	*595*	*623*	*654*	*604*	*635*
Bad-->Bad	*129*	*108*	*70*	*120*	*106*
Good-->Bad	*105*	*77*	*46*	*96*	*65*
Bad-->Good	*171*	*192*	*230*	*180*	*194*
Success Rate	*724*	*731*	*724*	*724*	*741*
Failure Rate	*276*	*269*	*276*	*276*	*259*
Success %	*72.4*	*73.1*	*72.4*	*72.4*	*74.1*
Failure %	*27.6*	*26.9*	*27.6*	*27.6*	*25.9*

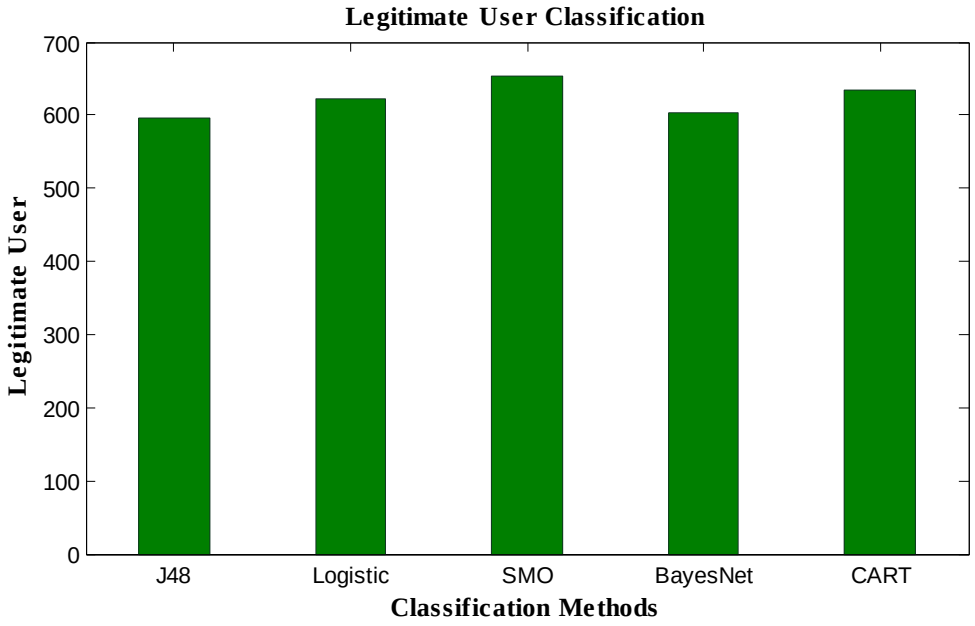

Figure 2. Legitimate User Classification

Figure 2 shows the classification of legitimate users using C5.0 using J48, Logistic Regression, SVM using SMO, BayesNet and CART. With a given credit card data set, SMO has classified 654 legitimate users as legitimate users. Next higher classification is given by the CART.

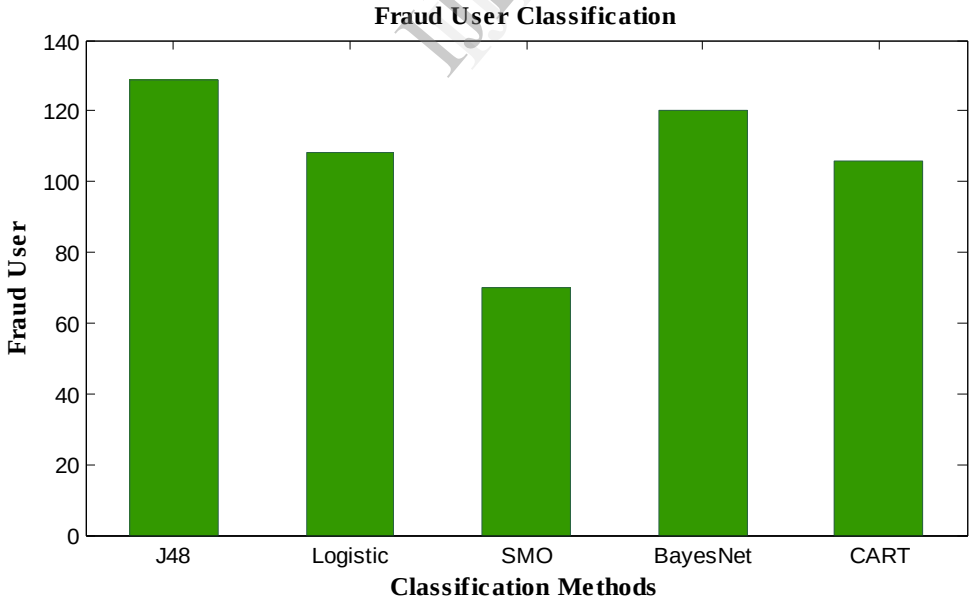

Figure 3. Fraud User Classification

Figure 3 shows the classification of fraud users using C5.0 using J48, Logistic Regression, SVM using SMO, BayesNet and CART. With a given credit card data set, J48

gives lesser classification of Bad → Good (Fraud user has been classified as Legitimate user). Next lesser classification is given by the BayesNet.

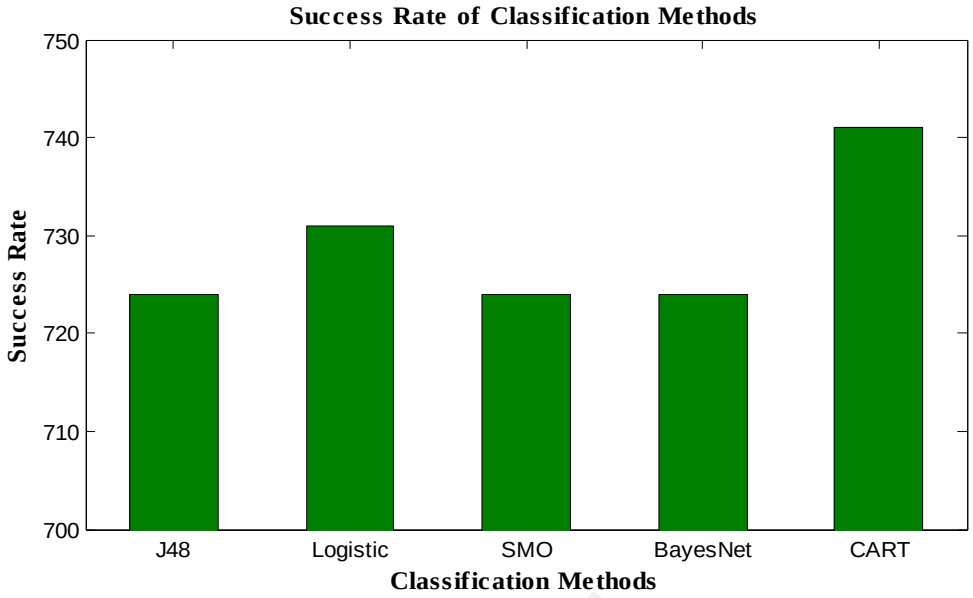

Figure 4. Success Rate of Classification Methods

Logistic Regression method is providing 73.1% success rate and CART gets the highest success rate 74.1%. Hence it is shown here that depending upon the success rate CART outperforms the other models whereas considering the Bad → Good classification J48 shows better performance.

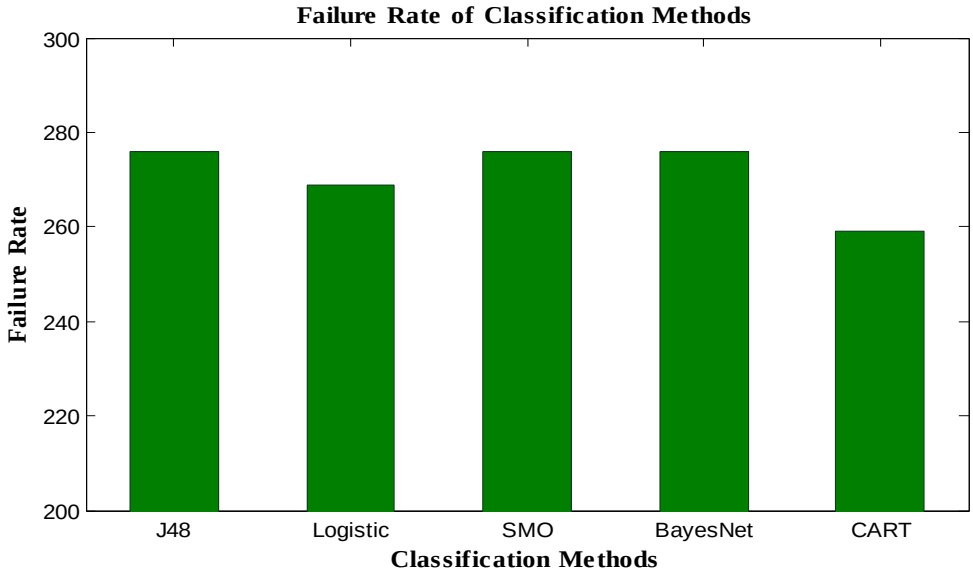

Figure 5. Failure Rate of Classification Methods

5. Conclusion

Revealing fraudulence in credit card approval is important for the efficient processing of credit applications. To improve security of the credit card approval systems in an automatic and effective way, building an accurate and efficient credit card approval system is one of the key tasks for the financial institutions. In this paper, five classification methods were used to detect fraud in credit card approval process in the banking sector. This work demonstrates the advantages of applying the data mining techniques including decision trees (C5.0 & CART), SVM using SMO, Logistic Regression and BayesNet to reveal fraudulence in credit card approval process. It reduces the financial institution's risk. Any financial institution needs to retain its customers. A legitimate customer must not be classified as Fraud or a Fraud must not be classified as Legitimate. Hence it is shown here that depending upon the success rate CART outperforms the other models whereas considering the Bad → Good classification J48 shows better performance. Hence It is concluded from this work that while making decisions on classifying customers combination of different classification models need to be used to make correct decision about a customer.

References

1) K. Chitra, B.Subashini, Customer Retention in Banking Sector using Predictive Data Mining Technique, International Conference on Information Technology, Alzaytoonah University, Amman, Jordan, www.zuj.edu.jo/conferences/icit11/paperlist/Papers/
2) K. Chitra, B.Subashini, Automatic Credit Approval using Classification Method, International Journal of Scientific & Engineering Research (IJSER), Volume 4, Issue 7, July-2013 2027 ISSN 2229-5518.
3) K. Chitra, B.Subashini, Fraud Detection in the Banking Sector, Proceedings of National Level Seminar on Globalization and its Emerging Trends, December 2012.
4) K. Chitra, B.Subashini, An Efficient Algorithm for Fraud Detecting Credit Card Frauds, Proceedings of State Level Seminar on Emerging Trends in Banking Industry, March 2013.
5) Richard J. Bolton and David J. Hand, "Unsupervised Profiling Methods for Fraud Detection", Technical Report (Department of Mathematics, Imperial College, London), 2002.
6) Mayur Shetty, "RBI moves to check credit card frauds", The Times of India, March 1st 2013.
7) Bharati M. Ramageri, "Data Mining Techniques and Applications", Indian Journal of Computer Science and Engineering, Vol. 1 No. 4 301-305.
8) Margaret H. Dunham, "Data Mining Introductory and Advanced Topics", Pearson Education, Sixth Impression, 2009.

9) B. C. da Rocha and R. T. de Sousa, "Identifying Bank Frauds using Crisp-Dm and Decision Trees", International journal of computer science & information Technology (IJCSIT) Vol.2, No.5, October 2010.
10) Classification: Naive Bayes vs Logistic Regression, John Halloran, University of Hawaii at Manoa EE 645, Fall 2009
11) Weka, Machine Learning Group at the University of Waikato.