# Automatic Credit Approval using Classification Method

Dr. K. Chitra, Mrs. B. Subashini

**Abstract** – This research paper aims to evaluate the performance and accuracy of classification models based on decision trees(C5.0 & CART), Support Vector Machine(SVM) and Logistic Regression with a dataset. Three methods to detect fraud are presented. Automatic credit approval is the most significant process in the banking sector and financial institutions. It prevents the fraud which is going to happen. So this paper proposes a good solution to the credit approval using the above methods.

**Index Terms** – Classification, Credit approval, Data Mining, Fraud, Logistic Regression, SVM

## 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 space of time[1]. 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[2]. So now a days, 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 compares the classification methods such as decision trees, Support Vector Machine(SVM) and Logistic Regression depends on the performance metric s such as performance and accuracy.

The rest of this paper is organized as follows: Section 2 describes the basics of data mining. Section 3 provides the classification method which are mainly used to apply in the credit approval dataset. Section 4 presents the performance metrics to compare the classification methods. Section 5 gives the details of experiments and results. Section 6 concludes this paper.

Dr. K. Chitra - Assistant Professor, Department of Computer Science, Government Arts College, Melur, Madurai.
Mrs. B. Subashini - Ph. D. Research Scholar, Manonmaniam Sundaranar University, Tirunelveli.
Assistant Professor, Department of Computer Science, V.V.Vanniaperumal College for Women, Virudhunagar.

## 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. 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[3].

### 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[4].

#### a 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[5].

### b 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.

### c 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.

### d 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[6].

## 3 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.

## 4 Experiments and Results

For the experimental work in this paper, it was not yet able to obtain a suitable real credit card approval dataset. Consequently, 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 originally provided by Quinlan in his studies and to induce classification models for assessing credit card applications. The dataset has 15 attribute plus the class label attribute. All attribute names and values were changed, before being released, to meaningless symbols to protect the confidentiality of the data. The dataset is summarized in Table 1.

The dataset is interesting because there is a good mix of attributes: continuous, nominal with small numbers of values, and nominal with large numbers of values. There are 690 instances in this dataset, with 307(44.5%) being positive(credit approved) and 383 (55.5%) being negative(credit denied).

**Table 1**

| Attribute | Type | Values |
|---|---|---|
| A1 | Nominal | a,b |
| A2 | Continuous | 13.75 – 80.25 |
| A3 | Continuous | 0 – 28 |
| A4 | Nominal | u, y, l, t |
| A5 | Nominal | g, p, gg |
| A6 | Nominal | c, d, cc, i, j, k, m, r, q, w, x, e, aa, ff |
| A7 | Nominal | v, hh, bb, j, n, z, dd, ff, o |

| | | |
|---|---|---|
| A8 | Continuous | 0 – 28.5 |
| A9 | Nominal | t, f |
| A10 | Nominal | t. f |
| A11 | Continuous | 0 – 67 |
| A12 | Nominal | t, f |
| A13 | Nominal | g, p, s |
| A14 | Continuous | 0 -2000 |
| A15 | Continuous | 0 - 100000 |
| Class | Nominal | +, - |

The tests were made using the software Weka (Waikato Environment for Knowledge Analysis) which contain a 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. Weka 3.6.x and 3.7.x have extensive help facilities built in and come with a comprehensive manual[7].

For the application in a decision tree, the algorithm used was J48, which is an evolution of C5.0. Here, the training algorithm took only 0.02 seconds. The network showed a rate of 4.8309% errors and accuracy of 0.953 for the correct cases, where there is no fraud, as well as 0.950 for the cases where there is fraud. The false positives are 13, and the false negatives are 7 from the confusion matrix. The classifier output for C5.0 is follows.

```
=== Run information ===
Number of Leaves  :     43
Size of the tree  :     62
Time taken to build model: 0.02 seconds
=== Evaluation on training set ===
Time taken to test model on training data: 0 seconds
=== Summary ===
Correctly Classified Instances         394               95.1691 %
Incorrectly Classified Instances        20                4.8309 %
Kappa statistic                          0.9033
Mean absolute error                      0.0874
Root mean squared error                  0.208
Relative absolute error                 17.5234 %
Root relative squared error             41.6465 %
Coverage of cases (0.95 level)          99.2754 %
Mean rel. region size (0.95 level)      74.0338 %
Total Number of Instances              414
=== Detailed Accuracy By Class ===
TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
0.940    0.036    0.967      0.940   0.953      0.904    0.970     0.969     +
0.964    0.060    0.936      0.964   0.950      0.904    0.970     0.948     -
Weighted Avg.  0.952    0.047    0.952      0.952   0.952      0.904    0.970     0.959
=== Confusion Matrix ===
   a   b   <-- classified as
 204  13 |   a = +
   7 190 |   b = -
```

Likewise C5.0, the other chosen methods to build classifier models are CART from decision tree methods, SVMs with kernels of polynomial functions and Logistic Regression. All these methods are used to develop models using the data sets. The Table 2 summarizes the success rate and time taken to build model for all four classification methods.

**Table 2**

| Classification Method | Success Rate | Time taken to build model |
|---|---|---|
| C5.0 | 95.1691 | 0.02 |
| CART | 84.058 | 0.19 |
| SVM | 84.7826 | 0.31 |
| Logistic Regression | 87.9227 | 0.08 |

## 5 Conclusion

Automatic credit approval is important for the efficient processing of credit applications. To improve security of the credit approval systems in an automatic and effective way, building an accurate and efficient credit approval system is one of the key tasks for the financial institutions. In this study, four classification methods were used to build fraud detecting models. The work demonstrates the advantages of applying the data mining techniques including decision trees, SVM and Logistic Regression to the automatic credit approval problem for the purpose of reducing the bank's risk. The results show that the proposed classifiers of CART outperform other approaches in solving the problem under investigation.

## References

[1] Richard J. Bolton and David J. Hand, "Unsupervised Profiling Methods for Fraud Detection", Technical Report (Department of Mathematics, Imperial College, London), 2002.

[2] Mayur Shetty, "RBI moves to check credit card frauds", The Times of India, March 1st 2013.

[3] Bharati M. Ramageri, "Data Mining Techniques and Applications", Indian Journal of Computer Science and Engineering, Vol. 1 No. 4 301-305.

[4] Margaret H. Dunham, "Data Mining Introductory and Advanced Topics", Pearson Education, Sixth Impression, 2009.

[5] 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.
[6] Classification: Naive Bayes vs Logistic Regression, John Halloran, University of Hawaii at Manoa EE 645, Fall 2009
[7] Weka, Machine Learning Group at the University of Waikato