Comparing decision trees with logistic regression for credit risk analysis

Article · January 2006

CITATIONS
28

READS
6,682

2 authors:

S. S. Satchidananda
Chidambara Research Karnataka India
8 PUBLICATIONS **31** CITATIONS
SEE PROFILE

J. B. Simha
Reva University
53 PUBLICATIONS **285** CITATIONS
SEE PROFILE

Comparing decision trees with logistic regression for credit risk analysis

S S Satchidananda
Research Director & Professor, CBIT,
International Institute of Information Technology, Bangalore, India
sssatchidananda@gmail.com

Jay B.Simha
Abiba Systems
Bangalore, India
jay.b.simha@abibasystems.com

Abstract

Credit risk evaluation is an important and interesting problem in financial analysis domain. Several techniques like expert systems, neural networks etc. have been used for credit rating. However these methods have limitations of knowledge bottleneck, slow learning etc. Recently decision trees have been proposed as the white-box models for learning and classification. In this work an attempt has been made to evaluate decision tree learning scheme with a logistic regression classifier on default risk of agricultural loans. It has been found that the decision tree classifiers will produce good results with parsimonious models.

1. Introduction

Credit-risk evaluation is a very challenging and important data mining problem in the domain of financial analysis. Many classification methods have been suggested in the literature to tackle this problem. But most of them are not accepted by the practicing experts due to various reasons. Against this background, we examine two classifiers in terms of their accuracy, True Positive and False Negatives with a view to evaluating their comparative efficacy. We use the data from two banks in India pertaining to the agricultural production loans given to farmers in and around Honavar, a backward block in Karnataka, India.

2. Literature survey

Artificial intelligence technologies have been employed for the development of credit scoring software systems that can meet the emerging needs and requirements [4,5]. On the one hand, expert systems have the advantage of representing and reasoning about relations amongst symbolic objects. This facilitates the task of generating explanations about objects and about inferences on the relations amongst objects. The disadvantage of expert systems is that the relations embedded in their knowledge base are pre-defined and their maintenance can become a tedious task. The increasing complexity of credit instruments, the volatility of the economic conditions and the importance of risk management in minimizing losses of credit portfolios impose the need for decision support systems with learning capabilities for dynamically analyzing various sources of

historical data and capturing complex relations amongst the most important attributes for credit evaluation.

Neural networks, especially, have received a lot of attention in credit scoring because of their universal approximation property. However, a major drawback associated with the use of neural networks for decision-making is their lack of explanation capability. While they can achieve a high predictive accuracy rate, the reasoning behind how they reach their decisions is not readily available [1].

Logistic regression

Logistic regression is a variation of ordinary regression which is used when the dependent variable is a binary variable (i. e., it takes only two values, which usually represent the occurrence or non-occurrence of some outcome event) and the independent (input) variables are continuous, categorical, or both. Unlike ordinary linear regression, logistic regression does not assume that the relationship between the independent variables and the dependent variable is a linear one. Nor does it assume that the dependent variable or the error terms are distributed normally.

The form of the model is

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

where p is the probability that Y=1 and X_1, X_2,.. .,X_k are the independent variables (predictors). β_0 , β_1, β_2, ... β_k are known as the regression coefficients, which have to be estimated from the data. Logistic regression estimates the probability of a certain event occurring.

Logistic regression, thus, forms a predictor variable (log (p/(1-p)) which is a linear combination of the explanatory variables. The values of this predictor variable are then transformed into probabilities by a logistic function. This has been widely used in credit scoring applications due to its simplicity and explainability.

Decision trees

A decision tree or a rule based classifier is a predictive model; that is, a mapping of observations about an item to conclusions about the item's target value. Each interior node corresponds to a variable; an arc to a child represents a possible value of that variable. A leaf represents the predicted value of target variable given the values of the variables represented by the path from the root. In decision tree learning, a decision tree describes a tree structure wherein leaves represent classifications and branches represent conjunctions of features that lead to those classifications [2]. A decision tree can be learned by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner. The recursion is

completed when splitting is non-feasible or when a singular classification can be applied to each element of the derived subset.

C4.5 is an improved version of an earlier decision tree based learner called ID3, which it self was based on Hunt's [3] concept learning system. The attributes are tested for selection on several criteria like minimum entropy, information gain etc., this has become a popular classifier in data mining due to both flexibility of learning and comprehension of results.

Experiments

The data used in this study comes from two banks .For preparing the training and test sets, the data set is balanced for Positive and Negative cases with cluster prototypes built from negative samples using k –means cluster algorithm in which the number of clusters is set to the number of positive samples. This approach will reduce the information loss compared to balancing the sample with under sampling or over sampling.

Variables Used

The original dataset contained about 25 variables. However, after exploratory analysis, the following variables were identified.

- CROP: Crop for which the loan was taken
- PI: Procured inputs
- SIRR: Spent for irrigation
- PWA: Purchased work animals
- PAI: Purchased agricultural implements
- RWC: repaired wells/canals
- LDW: Land development works done
- RDH: Repaired dwelling house
- SMP: Spent for miscellaneous purposes
- CDP: Crops damaged by pests
- CDNC: Crops damaged by natural calamities/causes
- LMP: Low market price
- RA: Repaid advance money
- SUE: Spent on unexpected expenditure
- IFA: Income

- PRO: Land
- CAR: Cleared and renewed*
- FS: Family size
- RISK: Risk class (Low or High)

The data prepared in the previous step has been analysed using decision tree and the results are compared with that of the logistic regression. The data was analysed with ten fold cross validation. The twelve rules induced from the decision tree have been shown in fig 1. Similarly the logistic regression model fitted to the data is shown in fig 2.

Rule No.	Rule	Class Name	Confidence (%)
1	X4 = N AND X16 <= 3 AND X14 = N	Y	100
2	X4 = N AND X16 <= 5 AND X14 = N AND X16 > 3 AND X15 = N	N	80.357
3	X4 = N AND X16 <= 5 AND X14 = N AND X16 > 3 AND X15 = Y AND X2 = N	Y	100
4	X4 = N AND X16 <= 5 AND X14 = N AND X16 > 3 AND X15 = Y AND X2 = Y AND X3 = N	N	84.906
5	X4 = N AND X16 <= 5 AND X14 = N AND X16 > 3 AND X15 = Y AND X2 = Y AND X3 = Y	Y	100
6	X4 = N AND X16 <= 5 AND X14 = Y	Y	91.667
7	X4 = N AND X16 > 5 AND X8 = N AND X2 = N AND X3 = N	N	100
8	X4 = N AND X16 > 5 AND X8 = N AND X2 = N AND X3 = Y	Y	100
9	X4 = N AND X16 > 5 AND X8 = N AND X2 = Y	N	100
10	X4 = N AND X16 > 5 AND X8 = Y	Y	66.667
11	X4 = Y	Y	83.333

Fig 1.Rules induced from Decision tree model

Variable Name	Coefficient Estimate
Intercept	-9.555
X2	0.289
X3	-3.14
X4	-7.138
X5	-0.402
X6	-1.462
X7	4.59
X8	-1.278
X9	-1.092
X10	3.318
X11	2.73
X12	-1.712
X13	1.977
X14	-1.616
X15	-2.184
X16	1.725
X17	-0
X18	-2.081

Fig 2.Logistic regression model

The important variables identified in the model are shown in fig 3.

Model 1 (DT)	Model 2 (LR)
X2	X3
X3	X4
X4	X7
X14	X10
X15	X15
X16	X18

Fig 3.Important variables

Results and discussions

The models were tested for their effectiveness on several measures – percentage correctly classified (PCC), true positive percentage (TP), false positive percentage (FP) and precision. These results are shown in fig 4. The lift charts of the models are shown in fig 5 and fig 6.

Defaulters				
	PCC(O)	*TP (%)*	*FP (%)*	*Precision (%)*
Model 1 (DT)	92	93	10	90
Model 2 (LR)	83	83	16	83
Non defaulters				
	PCC(O)	*TP (%)*	*FP (%)*	*Precision (%)*
Model 1 (DT)	92	90	6	93
Model 2 (LR)	83	84	17	84

Fig 4.Results of classification

It can be observed that the decision tree model outperformed the logistic regression model on all the parameters. The lift chart also shows that the decision tree slightly outperforms the logistic regression in finding the positives. We argue that the reason for this performance is that the logistic regression is good at optimizing a linear hyper plane for binary classification. Since the data used is almost discrete in nature (except the income which was numeric), decision tree which works well with axis parallel surface generalization than logistic regression, which provides an axis oblique hyper surface. Further the explanation from the decision trees in terms of rules are more appealing to the domain experts and required less cognitive load compared to getting intuition from the logistic regression coefficients.

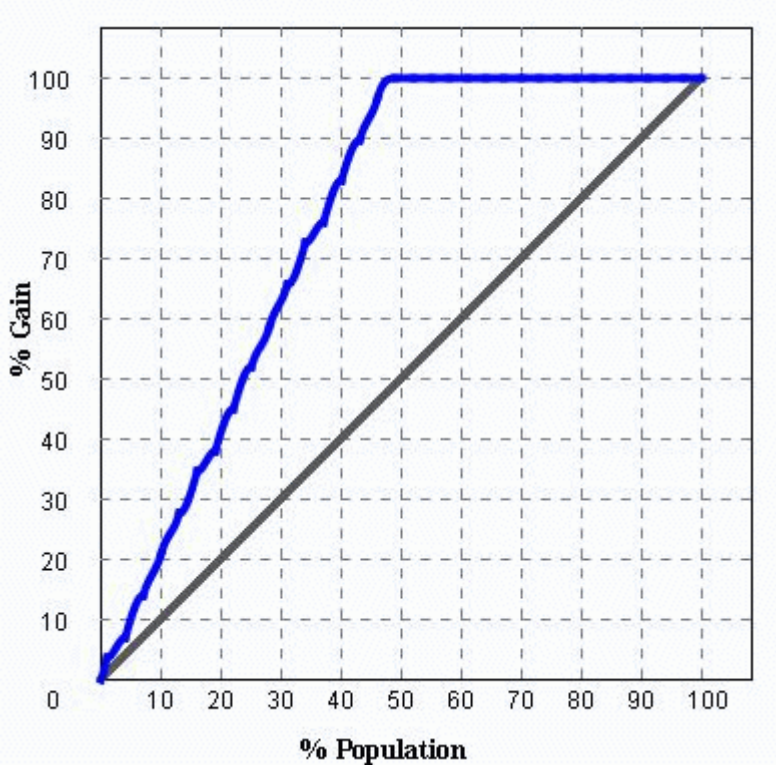

Fig 5.Lift chart for decision tree

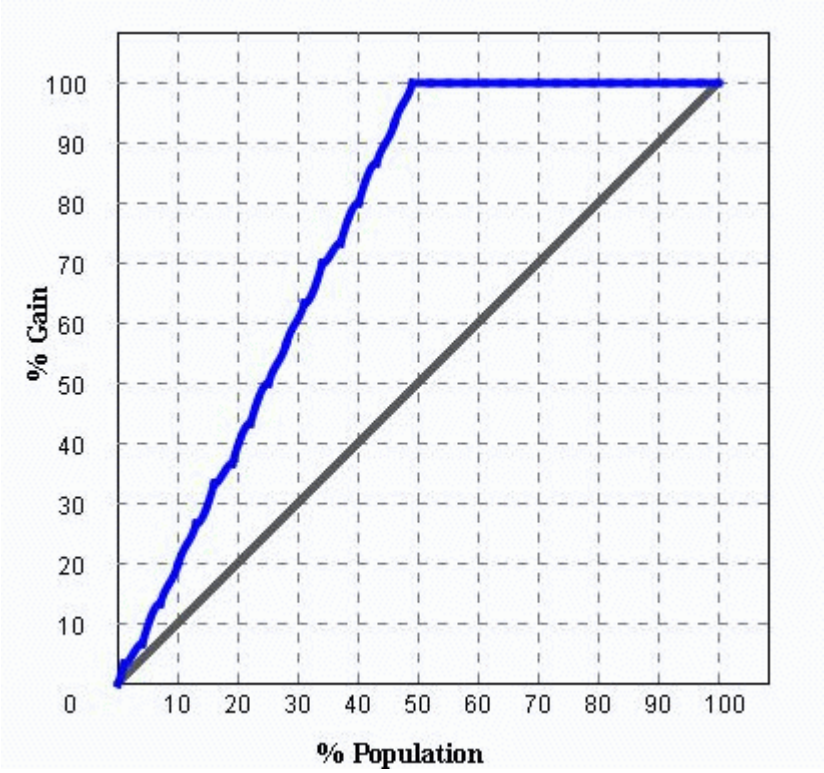

Fig 6.Lift chart for decision tree

Conclusions

In this paper, we have evaluated and contrasted decision tree classifiers with logistic regression classifiers for credit scoring. The evaluation was done by looking at the performance in terms of classification accuracy and the complexity of the trained classifiers. It was found that, the decision tree classifiers have a good performance and by using the proposed approach parsimonious and powerful models for financial credit scoring can be obtained. Work is under progress to investigate performance of proposed approach in comparison with other classification techniques for credit scoring.

References

[1] Baesens B., Setiono R., Mues C, Vanthienen J**.,** **"** Using Neural Network Rule Extraction and Decision Tables for Credit-Risk Evaluation", Management Science, Volume 49 , Issue 3, March 2003, pp:312 – 329.

[2] Breiman L, Friedman J, Olshen R.A and Stone C.J, 1984, "Classification and Regression Trees". Wadsworth.

[3] Hunt E.B., 1962, "Concept Learning", New York, Wiley

[4] Keyes, J. "Winning Back Investor's Confidence. Information Strategy", The Executive's Journal, Vol.1, 1992, pp: 42-44.

[5] Srinivasan, V., Kim, Y.H. "Credit Granting: A Comparative Analysis of Classification Procedures.", Journal of Finance. Vol. XLII, 1987, 665- 681.