An Application of Classification Models in Credit Risk Analysis

Ruan Ling-ying

Abstract A default risk is defined as the possibility that a borrower will not be able to pay back the principle or interest associated with a lending. Credit card business has high risk of delinquency as there is no collateral required before borrowing the money. Lenders usually collect a lot of information to learn the consumer risks. A conventional method to this problem is to examine combinations of the information variables that are likely to have influence. However, hunch can leave out important variables without being noticed. In this article, we introduce statistical models to conveniently predict the default risk based on an application to a real data of credit card business. Several potential improvements are also discussed.

Keywords Credit risk • Classification models • Logistic regression • Boosting • Random forests

1 Introduction

Credit risk analysis (finance risk analysis, loan default risk analysis) and credit risk management is important to financial institutions which provide loans to businesses and individuals. Credit loans and finances have risk of being defaulted. A default risk is defined as the possibility that a borrower will not be able to pay back the principle or interest. The biggest default risk involves unsecured lines of credit, such as credit cards. With an unsecured line of credit, it may be impossible for the lender to get back much of the investment, in case of a default. Credit cards and other unsecured lines of credit therefore carry a bigger risk when a default does occur. For this reason, credit providers must understand risk levels of credit.

R. Ling-ying (✉)
Chongqing Three Gorges University, Wanzhou, Chongqing, China
e-mail: ruanlingying@163.com

B. Xu (ed.), *2012 International Conference on Information Technology and Management Science (ICITMS 2012) Proceedings*, DOI 10.1007/978-3-642-34910-2_45,

Personal credit scores are normally the most critical information lenders need to obtain. They are provided by external credit bureaus and ratings agencies. Credit scores may indicate personal financial history and current situation. However, it does not tell you exactly what constitutes a “good” score from a “bad” score. More specifically, it does not tell you the level of risk for the lending you may be considering.

Except credit scores, credit providers often collect a vast amount of information on credit users. So credit risk profiling (finance risk profiling) is very important. The Pareto principle suggests that 80–90% of the credit defaults may come from 10 to 20% of the lending segments. Profiling the segments can reveal useful information for credit risk management. Information on credit users (or borrowers) often consists of dozens or even hundreds of variables, involving both categorical and numerical data with noisy information. Profiling is to identify factors or variables that best summarize the segments. Analyzing such vast information is an extremely difficult and challenging task. As the total number of variables increases, the number of combinations to be examined in this way grows exponentially. When a large number of variables involved, the number of combinations is too large to be examined manually. Thorough systematic accurate analysis is all but impossible.

Fortunately, this problem can be overcome with the methodology described here. Predictive modeling is an excellent technique for credit risk management. Predictive models are developed from past historical records of credit and consumer behavior. From the past information, predictive models can learn patterns of different credit default ratios, and can be used to predict risk levels of future credit loans. In this paper, we present several methods of predictive modeling and illustrate their performance based on a real data task. In Sect. 2, we briefly introduce the models and we apply them in the real task of credit card default risk classification in Sect. 3. And we conclude the findings in Sect. 4.

2 Methodology

Logistic regression (Hastie et al. 2009) is a popular linear model in classification to model the posterior probabilities of the K classes based on linear functions in the features x. The model has the form

$$\log \frac{P(G = j|X = x)}{P(G = K|X = x)} = \beta_{j0} + \beta_j^T x, \qquad j = 1, \cdots, K - 1 \tag{1}$$

It is specified in K-1 log-odds or logit transformations to satisfy the constraint that the probabilities sum to one. The choice of the denominator is arbitrary and the estimates are equivalent under the choice. Here the K-th class is chosen as the baseline. It is solved by iteratively reweighted least square (IRLS). For typical two-class problems, it is simplified and the posterior probabilities have the form

$$P\left(G=1|X=x\right)=\frac{\exp\left(\beta_{10}+\beta_1^T x\right)}{1+\exp\left(\beta_{10}+\beta_1^T x\right)},$$

$$P\left(G=2|X=x\right)=\frac{1}{1+\exp\left(\beta_{10}+\beta_1^T x\right)} \tag{2}$$

Random forests (Breiman 2001) select a subset of variables for each node and choosing one best split from the subset to build one tree based on bootstrap samples (Efron 1979). Many trees are grown iteratively and it returns the probability by voting. One promising property of random forest is that it can handle high dimensional data and it does not over-fit. It is robust to outliers and it's easy to implement because there are not many parameters to be tuned. The only tuning parameters for random forest are the total number of trees to grow and the size of subset to split at each node. However, random forest is not sensitive to the size of subset in a wide range actually. Several choices of the parameter is tried to get the best random forest model.

Boosting was first proposed in Schapirer (1990) and was received much attention since then (Freund 1995; Freund and Schapirer 1997; Schapire et al. 1998; Friedman et al. 2000). The idea of boosting is to generate weak learners iteratively and combine them as a committee to do the classification. Each weak learner is trained sequentially to perform better than random guess. The iteration goes with down weighting the correctly classified samples and increasing the weight for those misclassified samples to get the next weak learner. Each weak learner has a weight to vote which depends on its performance, i.e., the less error rate, the more weight. Classification is the weighted vote of these weak learners. There are many boosting algorithms. AdaBoostM1 is most commonly used. In this study, we tried different boosting algorithms including AdaBoostM1, LogitBoost and BlackBoost. Also, it is flexible to choose the type of weak learner. Decision stump is used as the base learner when using AdaBoostM1 and LogitBoost. Decision tree is the base learner when using BlackBoost. See Schapire et al. (1998) and Friedman et al. (2000) for more discussion of boosting.

3 Application

Credit card provides provide as many variables as possible to understand the risk. Starting with the raw variables, we discussed several ways to extract features (Ruan 2010). With all the possible features ready, stepwise logistic regression chooses ten variables. Compared with the logistic regression model with the raw variables only, i.e., out feature extraction, the AUC of the new logistic regression is lifted by 10.1% on in-sample validation and 9.7% on out-of-time validation. The kolmogorov smirnov (KS) statistic is lifted by 12.3% on in-sample validation and 11.8% on out-of-time validation. Figure 1a is the ROC on the out-of-time validation data set.

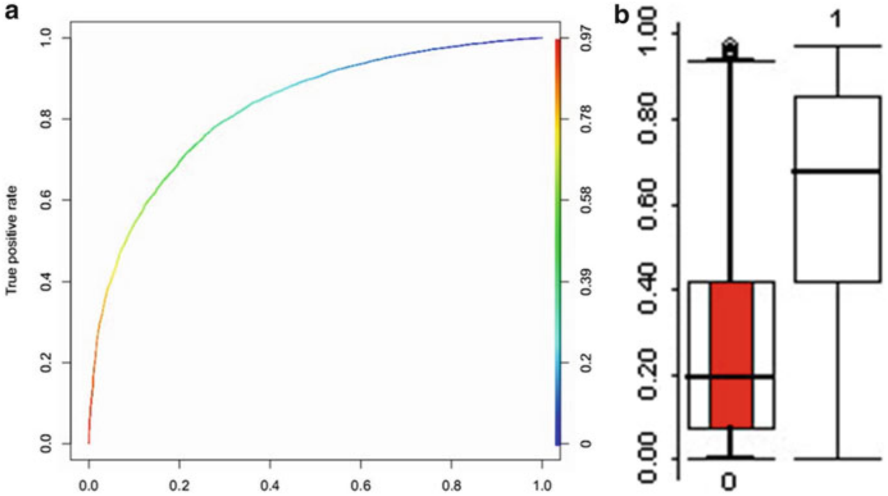

Fig. 1 (**a**) ROC on the out-of-time validation data set with AUC 92.6%; (**b**) *box plot* for the probability of good among the two classes on out-of-time validation, where 0 (*red*) represents bad and 1 (*white*) represents good

And Fig. 1b is the box plot for the probability of good among the two classes on out-of-time validation, where 0 (red) represents bad and 1 (white) represents good.

In random forest, all variables are first put into the model with subset size of 14 and 500 trees, which is decided by cross validation. By bagging (Breiman 1996), the importance of each variable available and is assessed by mean decrease in accuracy and mean decrease in node impurity. The previous one measures how one variable helps others and the latter one measures the individual contribution it has. To get a sparse model, we select part of the variables according to the importance order. For example, the geometric mean of the two metrics can be used as a criterion to select variables. Figure 2 is the importance picture.

For boosting, 500 iterations of AdaBoostM1 returns the best AdaBoostM1 model and 110 iterations of LogitBoost is the best one. To look into the importance of each variable, we check the weight of each decision stump. Table 1 lists the first ten weak classifiers in AdaBoostM1 and the corresponding weight and error rate. The higher the weight, the more important the variable is. And the more times one variable is selected as the split variable, the more important it is. For example, U_alpha is firstly selected as the split variable and the corresponding error rate and weight is 0.2809 and 0.94 respectively. U_alpha is the intercept from the regression, which means it is similar and highly correlated to current utilization. It is more than current utilization because it contains the information with the current utilization and the utilization trend over time. The performance of the two types of Boosting is similar to logistic regression.

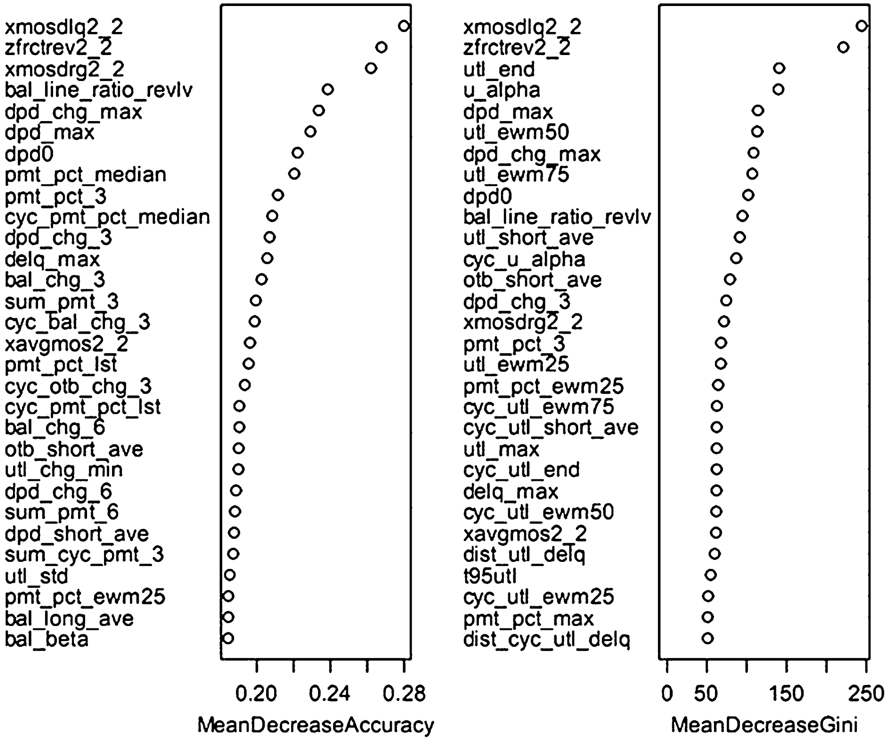

Fig. 2 Importance measures of variables in random forests. The *left one* is mean decrease in accuracy and the *firth one* is the mean decrease in node impurity

Table 1 The first ten weak classifiers in ADABOOSTM1

	Variable	Split	<=Split	>Split	Weight	Error rate
1	U_alpha	0.8232	Good	Bad	0.94	0.2809
2	Xmosdlq	1.5	Bad	Good	0.9	0.28905
3	Cyc_pmt_pct_3	0.114	Bad	Good	0.45	0.389361
4	zfrctrev	58.5	Good	Bad	0.29	0.428004
5	Dep_tot_am	832.74	Bad	Good	0.37	0.408541
6	U_alpha	0.8215	Good	Bad	0.29	0.428004
7	Cyc_ca_ewm25	7.9248	Good	Bad	0.2	0.450166
8	xavgmos	65.5	Bad	Good	0.23	0.442752
9	xmosdrg	8.5	Bad	Good	0.21	0.447692
10	Dpd_short_ave	2.166	Good	Bad	0.15	0.46257

4 Conclusion

This paper has empirically investigated the application of three classification models in the application of credit risk analysis. The data is based on the feature extraction illustrated in Ruan 2010. The Logistic regression shows better separation than the old one without feature extraction, which shows great advantage of doing feature extraction. The logistic regression model is a simple parsimonious model and it is easy to interpret. Random forests and boosting are suitable to handle a large set of variables.

References

Breiman L (1996) Bagging predictors. Mach Learn 24:123–140

Breiman L (2001) Random forests. Mach Learn 45:5–32

Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Stat 7(1):1–26

Freund Y (1995) Boosting a weak learning algorithm by majority. Inf Comput 121:256–285

Freund Y, Schapirer E (1997) A decision-theoretic generalization of online learning and an application to boosting. J Comput Syst Sci 55:119–139

Friedman J, Hastie T, Tibshirani R (2000) Additive logistic regression: a statistical view of boosting (with discussion). Ann Stat 28:337–407

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. Springer, New York

Ruan L (2010) An empirical study of feature extraction in the analysis of credit card risk. Technical report. Chongqing Three Gorges University

Schapire R, Freund Y, Bartlett P, Lee WS (1998) Boosting the margin: a new explanation for the effectiveness of voting methods. Ann Stat 26(5):1651–1686

Schapirer E (1990) The strength of weak learn ability. Mach Learn 5:197–227