Classification of Bad Accounts in Credit Card Industry

Chengwei Yuan

December 12, 2014

Introduction

Risk management is critical for a credit card company to survive in such competing industry. In addition to operational expenses, provisional loss is a major driver to a company's expense. The provisional loss arises due to the "bad" accounts booked – bank lends the money to customers who eventually do not have capability to pay back. In the risk management, there are generally two stages a company can take to manage and control credit risks. The first stage occurs when booking a customer. An aggressive underwriting strategy could book and approve high risk population who seeks for credit card; while a conservative policy may only focus on upmarket and affluent population. As expected, the first strategy could generate both high revenues (interests charged) and high expenses due to bad accounts booked, resulting in trivial incremental net income; and the second strategy could generate not only low revenues but also low losses, resulting in incremental net income be trivial as well. There is always trade-off for different strategies in terms revenue generation and loss control. Finding an optimal strategy is often difficult and needs to be adjusted accordingly due to internal or external factors such as macro-economic change, for example, almost all credit companies suppressed their approval rates for high risk segments of population and incurred huge financial loss during 2008/2009 economic depression. The second stage happens in customer management after the customer is booked. Although booked customers pass the first screen of risk control, the chance of false negative (false "good" accounts) could still be high. However, in the second stage, by leveraging their performance such as credit card utilization, payment information, risks can further be managed to control provisional loss. In our project, we will focus on the second stage of risk management, and particularly are interested in classifying if a booked account will be a "bad" account within 12 months since booked. Since an internal classification model is already available, our second interest is to train a better classifier to outperform the benchmark model.

Dataset

Data are not publicly available. There are three sources for both training and test samples: credit bureau data (from one of the largest three bureaus TransUnion, Experian or Equifax), consumer purchase behavior data (internally summarized purchase information) and customer experience data (information about how customers use digital like website or mobile on their accounts). The credit bureau and customer experience data contains 3 statements (from current snapshot statement) of historical observation, and consumer purchase behavior data have summarized 12 statements historical observation. There are three vintage snapshots for training sample: August '11 (40K accounts), March'12 (40K accounts) and June '12 (20K accounts). For the test sample, the snapshot comes from more recent vintages on January (10K) and March (10K) '13 year data. The target variable is dichotomous with value 1 or 0 indicating if an account is bad or not. The "bad" definition indicates that an account is either delinquent or charged-off in the future 12 months from current snapshot statement. The benchmark score (in the format of probability score) is also provided. In each data source, the unique account ID serves as the key index for combing all datasets.

Features and Preprocessing

Features

There are about 120 features per snapshot in the credit bureau data, resulting in 120 x 3 = 360 features; and 60 features per snapshot in the customer experience data, resulting in 60 x 3 = 180 features; and 300 features in the consumer purchase behavior data. Therefore, there are a total of 840 features in the raw training/test samples.

Preprocessing

Features selection

Since there are many available features in our samples, we decide to perform variables selection first before we train the models. As there are three sources of dataset, we choose to conduct variable selection separately for each of them. For credit bureau data, we only choose the current snapshot features (120 features) in variable selection while creating additional *trend* features as described in the next section. For customer experience data, we summarize the statement-wise features by either taking the max(indicator features) or mean(continuous features), thus reducing the features from 180 to 60 before the variable selection processing. The formula below gives the definition of customer experience data summarization:

$$Z = \begin{cases} \max(X_0, X_{-1}, X_{-2}), & if\ X_i s\ are\ indicators \\ \text{avg}(X_0, X_{-1}, X_{-2}), & if\ X_i s\ are\ continuous. \end{cases}$$

where X_0 represents the current snapshot feature, and X_{-1} and X_{-2} indicate the last two statement features.

We use the Treenet function in software SPM v7.0 (Salford Systems, San Diego CA) to perform quick variable selection. We also create an additional random variable (uniformly distributed) in the training sample and include it in variable selection. In the Treenet output, the variables are ranked by their predictive importance to the target. We apply two rules to determine the variables selected: first, for credit bureau variables, we aim to select at most top 50 variables; while for purchase behavior and customer experience data, we aim to select at most top 25 variables. Based on our understandings, bureau variables generally have stronger predictive power to risks than the other two types of data, thus we decide to keep more variables for bureau data; second, only variables that are ranked above the random variable will be kept in the selected variables list. We believe that any variables that are ranked under the random variable are more likely random noises. By applying rules above, we select a total of 81 features for model training purposes: 50 features from bureau data, 18 features from transaction and 13 features from digital data.

Additional features creation

As described above, for bureau data, we only choose the current snapshot features into the variable selection; however, we also think the trend of some features may have incremental values to predict risks, for example, the outstanding utilization trend (outstanding balance divided by credit line), the payment ratio trend (payment divided by last statement outstanding balance), the FICO trend and etc. We also create some indicators based on previous statements information such as if an account has made purchase or not, or if an account has made any payment or not and etc. As a result, we create a total of additional 12 features.

Therefore, there are a total of 93 features in our dataset for training purposes.

Feature treatment

We apply simple missing treatment and capping/flooring to features selected. If a feature has missing values, the missing observations will be imputed by medians; and the feature is capped by its 99th percentile and floored at 1st percentile.

Models

In this project, we test four classification machine learning models, which are implemented in SAS 9.3 (SAS Institute, Cary NC), Salford Predictive Modeler (SPM v7.0) and ***e1071*** package of the R tool (R Core Team, 2013): logistic regression, stochastic gradient boosting, random forests, and support vector machine (SVM).

The logistic regression (McCullagh and Nelder, 1989) is a popular method in banking industry, particularly in credit scorecard development (Thomas, 2002; Siddiqi, 2006). In a binary classification problem, the logistic regression models the $p(y = 1|x; \theta)$ as a sigmoid function of input features by $p(y = 1|x; \theta) = \frac{1}{1+\exp(-\theta^T x)}$. The classification can be determined by the decision boundary $p(y = 1|x) > C$, where C is a pre-specified threshold (generally $C = 0.5$). The logistic regression is easy for interpretation, but is usually limited by its constraints on capturing nonlinear relationships.

The gradient boosting method was proposed by Friedman (1999) and was later improved to stochastic gradient boosting by using the bagging procedure. The purpose of boosting methods is to sequentially construct a sequence of weak classifiers and then ensemble them through a weighted majority vote to produce the final prediction.

The random forests method was introduced by Breiman (2001). It is an ensemble method to improve the variance reduction of bagging by reducing the correlation between the trees, without increasing the variance too much. The de-correlation is achieved via the random selection of variables at nodes when growing trees. Random forests method has several advantages such as its capability of capturing the non-linear boundary between event and non-event, naturally embedded out-of-bag validation, no special treatment to input features, suitable for unbalanced data classification and etc.

The SVM (Cortes and Vapnik, 1995) classifier generally transforms the input attributes into a high dimensional feature space by introducing a mapping (linear or non-linear) via a kernel. When training the classifier, SVM only uses the related support vector points in feature space to find the optimal separating hyperplane. One popular kernel choice is the Gaussian kernel $K(x, z) = \exp\left(-\frac{||x-z||^2}{2\sigma^2}\right)$, which represents less parameters than other kernels. For a binary classification problem, the probability output can be generated by

$$p(y_j = 1|x) = \frac{1}{1 + \exp(\sum_{i=1}^{m} y_i \alpha_i K(x_i, x_j) + b)}.$$

Before fitting the SVM model, the input data is first standardized to a zero mean and one standard deviation. Also, we enable the *probability = TRUE* in the *svm()* function of *e1071* package to obtain probability type scores.

The model performance is compared by ROC and AUC (area under ROC curve - Bradley, 1997) on the test sample. The classifier that results in largest AUC will be treated as the best one. Our model will also be compared to the internal developed model to show the incremental improvement.

Results and Discussion

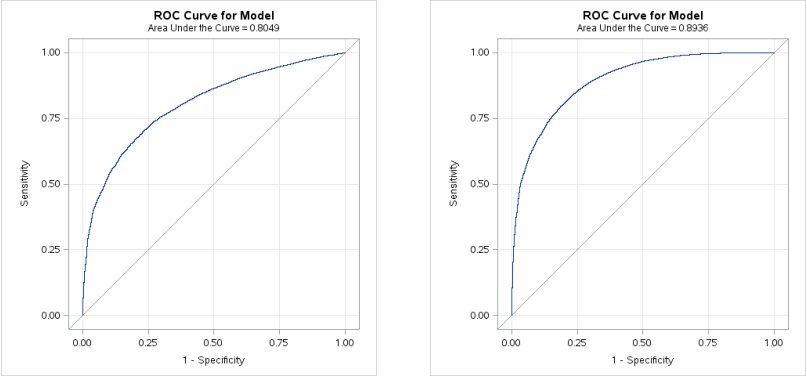
Results

We calculate areas under ROC curve on the test sample after four models are trained. The AUCs from benchmark and four models are provided in Table 1. Since random forests method generates the largest incremental values to the benchmark, we also provide the ROCs as shown in Figure 1.

Table 1: The AUCs of benchmark model and new models

Models	AUC		% increase of AUC to benchmark	RF vs other models
	Benchmark	New Model		
Logistic regression	0.805	0.836	3.85%	6.94%
SVM		0.868	7.83%	3.00%
Stochastic gradient boosting		0.882	9.57%	1.36%
Random forest (RF)		0.894	11.06%	-

Figure 1: ROC curves of benchmark and random forest models.
Left: benchmark; Right: random forest



Discussion

The obtained results for the four models are shown in Table 1 in terms of AUC metric computed on the test sample. The best result is achieved by the random forest model which outperforms logistic regression (improvement of 6.94%), SVM (improvement of 3.00%) and stochastic gradient boosting method (improvement of 1.36%), while all models can beat the benchmark model. It is expected that all models could beat the benchmark, since the benchmark model only applied the logistic regression on credit bureau data. The new logistic regression model has smallest incremental value in terms of AUC to the benchmark, while SVM, stochastic gradient boosting and random forest have substantial improvement to the benchmark.

We also check the top 10 variables that have most predictive power to forecast the likelihood to be charged-off. They follow into three categories: balance utilization, FICO and the carried

total/highest balance, all of which are business intuitive. For example, the higher utilization indicates the higher risk to be "bad" account; high FICO score means the low risk population, and carrying higher balance explains the customers tend to revolve their balance and are lack of capabilities to pay off and thus are more risky.

Although the sophisticated machine learning methods can beat the benchmark (logistic regression) in terms of model performance, it may give rise to a problem with respect to model implementation and interpretability. The credit card industry is highly regulated by OCC (Office of the Comptroller of the Currency), so companies are required to provide transparency of model structure and interpretation to regulators. However, random forest and stochastic gradient boosting, as ensemble methods, are quite complicated as trees are ensembled together and thus encounter difficulties for interpretation. In addition, their implementation into the internal production system could also be challenging.

Future

We have been able to apply machine learning techniques to train better models that can outperform the benchmark model, however, considering the difficulties of model interpretability and implementation, in the future, we can do more researches on:

1) Find creative solutions to provide enough transparency of ensemble methods to regulators and enable implementation in the production system.

2) Improve logistic regression under the framework of gradient boosting methods so that the trained model can have both improved model performance and clear interpretation.

Reference

[1] Bradley, P. A. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30(7): 1145-1159.

[2] Breiman, L. (2001). Random forest. *Machine Learning* 45 (1): 5-32.

[3] Cortes, C. and C. Vapnik. (1995). Support vector networks. *Machine Learning* 20 (3): 273-297.

[4] Friedman, J. H. (1999). Stochastic gradient boosting. Technical report, Dept. of Statistics, Stanford University.

[5] McCullagh, P. and J. A. Nelder. (1989). *Generalized linear models*. Second edition. London: Chapman and Hall.

[6] R Core Team. (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org/.

[7] Salford Systems. Available at http://www.salford-systems.com/products/spm.

[8] Siddiqi, N. (2006). *Credit risk scorecards: developing and implementing intelligent credit scoring*. John Wiley & Sons, Inc. Hoboken, NJ.

[9] Thomas, L. C., D. B. Edelman, and J. N. Crook. (2002). *Credit scoring and its applications*. SIAM. Philadelphia, PA.