Clustering and Prediction for Credit Line Optimization

Ira J. Haimowitz   Henry Schwarz

General Electric Corporate Research and Development
One Research Circle
Niskayuna, NY 12309
E-mail: {haimowitz, schwarz}@crd.ge.com

Abstract

Credit granting businesses face a challenging environment due to the wide variety of customer behaviors. While only some customers use their credit and pay regularly, a larger percentage may hardly use their available credit. As a key risk management issue, small percentages of customers become delinquent in their payments, and others become bankrupt, requiring write-off. As a business decides upon the deal structure (credit line, repayment terms, interest rate, etc.) of a customer, that business needs to optimize the deal structure considering the uncertainty of that customer's behavior.

We have developed a framework for credit customer optimization based on clustering and prediction. First customer clusters are formed by using hierarchical clustering from past credit performance data. Then, external data, as from a credit bureau, is used to predict the probabilities of membership for each performance cluster. The prediction is done using classification and regression trees (CART). We show an example of this framework used for initial credit line optimization.

KEYWORDS: risk management, optimization, clustering, prediction, decision tree induction.

Introduction

Managing New Customer Risk Under Uncertainty

Credit granting businesses regularly must decide for each new customer what the financing structure should be, including credit line, repayment terms, interest rate, etc. Examples of credit businesses facing these challenges are:

1. automobile lessors
2. mortgage banks
3. credit card issuers
4. retail merchants that extend credit

Credit businesses must make these decisions realizing that their customers fall into *widely different classes* from risk and profitability perspectives. While only some customers use their credit and pay regularly, a larger percentage may hardly use their available credit. Additionally, small percentages of customers will become delinquent, and others become bankrupt, requiring write-off. Despite the uncertainty, the credit business must somehow predict a new customer's future behavior when given credit, and determine an optimal deal financing structure for that customer.

Inadequacy of Traditional Scoring Models

The most typical data used for prediction of customer behavior is from *credit bureaus;* examples are Dun & Bradstreet and Equifax. Traditionally, consultants use external credit bureau data to develop *scoring models* that predict a binary variable such as: delinquency or not, or default or not. Scoring models take two complementary historical customer sets, the "good" and "bad" performers, and use credit bureau data to distinguish between the two sets. Typically logistic regression is used; a good survey can be read in [Rosenberg and Gleit]. Scoring models can be applied for new credit authorizations.

Scoring models inherently examine just one customer performance measure, and thus yield an incomplete picture of the behavior of a new credit customer. They are inadequate for answering more subtle questions about the customer like:

- How much of the credit line will the customer use?
- What percentage of the monthly statement will the customer pay, versus revolving (and thus generating service charge payments)?
- How profitable will this credit customer be for my financial company?

Additionally, traditional scoring models do not treat credit line as an endogenous, independent variable. In our work we have done so, and aim to optimize a new customer's expected long-term profitability as a function of the credit line.

Framework for credit customer optimization

Our framework is an extension of the traditional scoring model approach that captures more aspects of an expected customer's behavior, and can be used for optimizing a new customer's deal structure. The framework is illustrated in Figure 1, and consists of three phases:

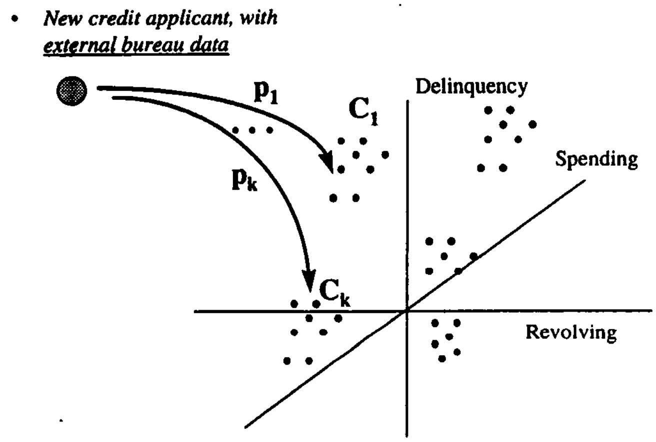

Figure 1: Clustering-based framework for optimizing a credit policy.

1. Clustering and Curve Fitting
2. Prediction of Cluster Probabilities
3. Optimization Model

In phase 1, historical accounts are clustered into K groups based on many months of performance data that account for the customers' patterns of spending and paying bills. The rationale for the clustering is to divide the historical accounts into their different behavior patterns. Preferably, each account should include monthly observations from the beginning of the account until a fairly long performance period. This allows delinquent and bankrupt accounts to reach those undesirable states. In practice, we have found 21 or more months as a suitable time horizon. Within each cluster, a curve is fit that maps the relationship between the deal structure variables and the ***expected net present value*** (NPV) to the credit company. NPV will be described more in section 3.

In phase 2, decision tree induction, in the form of regression trees, is used to predict the probability of cluster membership for new accounts. The decision tree induction uses old external credit bureau data for the historical accounts, using a snapshot of the time those accounts were applying for credit. The particular decision tree used is CART, for Classification and Regression Trees [Breiman et. al.], which is part of the S-Plus statistics package. The output of the decision tree is a set of rules, with each rule predicting K probabilities, p_1 to p_k, of membership in each of the K clusters.

In phase 3, the optimal deal structure for a new customer is determined by maximizing the overall expected net present value of a customer over all values of that deal structure. The overall net present value is computed as follows, say for optimizing a deal vector V:

$$\begin{aligned} \text{Exp(NPV} \mid \text{V)} = &\ \text{Pr(cluster 1)} * \text{Exp(NPV} \mid \text{cluster 1, V)} \\ &+ \text{Pr(cluster 2)} * \text{Exp(NPV} \mid \text{cluster 2, V)} \\ &+ \ldots \\ &+ \text{Pr(cluster K)} * \text{Exp(NPV} \mid \text{cluster K,V)} \end{aligned}$$

Where Exp(NPV | V) is the expected dollar net present value for a customer with deal vector (including credit line) V.

While there is no guarantee that this overall expected net present value will have a unique maximum, we have found unique optima in practice for optimizing against the one deal variable of initial credit line. That application is the subject of the rest of this paper.

Application: Credit Line Optimization

We have applied this modeling framework to optimizing the initial credit lines for new customers, a typical project within financial risk management of credit companies. Credit line assignment is a risk management issue for two primary reasons:

1. Customers that write-off tend to do so close to their credit limit
2. Unused credit line is excess "exposure" for a credit company, which is highly discouraged because a customer may in hard financial times become risky.

To protect business confidentiality, we describe the main qualitative results while omitting the specific data attributes use and the financial dollar amounts. We divided our roughly 82,000 accounts into training and validation sets based on the time of initial credit application. The 55,000 accounts applying for credit in October and November 1993 were the training set; the 27,000 applying for credit in December 1993 were the holdout set. This experimental design let us test the model's predictive ability.

Cluster Descriptions

The attributes used in clustering the credit customers were related to:

- Spending patterns over 21 months
- Patterns of paying monthly bills.
- Usage of the credit line.
- Other risk related attributes.

First a random 10% sample was taken, and hierarchical clustering performed (which has run time of $O(N^2)$ for N accounts). The optimal number of clusters was determined from distance changes in the resulting dendrogram. After comparing results with two random samples, we determined that 5 clusters was best. Then, we used the five cluster centers as inputs in a K-means (or iterative nearest neighbor) clustering (which has run time of O(N)) on all of the 55,000 observations. K-means was run for 1,000 iterations, or until convergence.

The five clusters are listed here with their general characteristics:

1. Usually on time with payments, pay most of their monthly balance, use some of their credit line, fairly high sales, and fairly profitable.
2. Fairly delinquent accounts, pay some of their monthly balance, high sales, and very profitable. Should be treated with caution in times of recession.
3. On time with payments, but very little sales activity. Not very profitable.
4. Very delinquent; all of these are write-offs. Generate fairly high sales but are unprofitable. Creditors lose money on these.
5. Mixture of on-time and delinquent accounts, generate high sales, and are very profitable, especially at lower credit lines.

Net Present Value Relationships Within Clusters

The *net present value* (NPV) of an investment is defined as the net income that investment generates, with future income discounted to the time of the original investment. Net present value is often used by corporations in budgeting capital investments [Brealy and Myers], and is recommended as a good quantitative measure of the success of a targeted marketing campaign [Hughes]. NPV is also a natural measure of the profitability of an individual customer, such as a catalogue recipient [Bitran and Mondschein]. Another example is a long-term credit customer, because that customer's monthly payments are likely to last over a period of several years.

An NPV calculation is dependent on the financial application, and generally includes both expenses and revenue. Expenses include the cost of acquiring the account, cost of mailing bills, and written-off dollars from unpaid balances. Revenue includes payments of bills received and service charges received.

In this example, we have examined the relationship between initial credit line and NPV for accounts within each of the five clusters. The relative relationships are illustrated in Figure 2. Each graph was determined by plotting the truncated mean NPV for the accounts receiving that initial credit line. The plots were then smoothed by Loess curve fitting. Clusters 1 and 2 are more valuable with increased credit lines, with a plateau at higher lines. Cluster 3 shows little profitability for any credit line. Cluster 4, consisting entirely of write-off accounts, shows negative and decreasing NPV for increasing credit lines. Cluster 5 increases in profitability for lower accounts, then decreases for higher accounts as the write-off and delinquent effect overcomes the profitable effect. As can be seen from these bivariate relationships, modeling at the individual cluster level can be more accurate than at the overall population level.

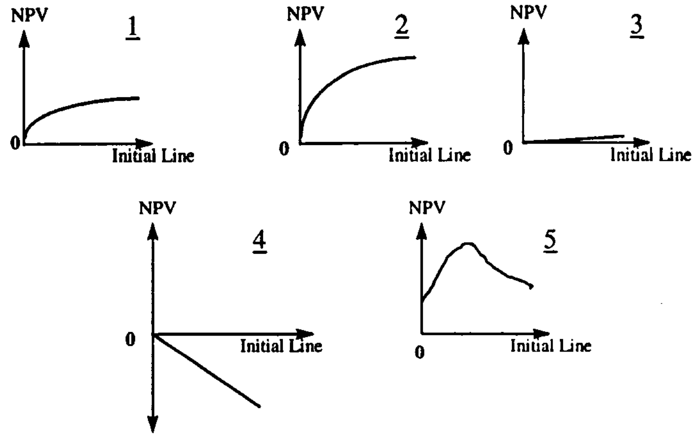

Figure 2: Models of Net Present Value versus initial credit line for five historical clusters.

CART Rule Results

The CART analysis linked external credit bureau data from the time of credit application to the cluster numbers for all 55,000 accounts in the training set. Thus the decision tree induction predicted the probability of cluster membership for credit accounts in this population. The CART analysis produced 17 rules, with probabilities summarized in the table below. For example, the first rule says that if various credit variables are above or below particular thresholds then the probabilities of membership in the 5 clusters are: (0.24690 0.4599 0.2229 0.043050 0.027250). Probabilities are estimated as the frequencies of membership in each of the 5 clusters, divided by number of observations meeting that rule's criteria.

Note in particular that rules 5 and 7 have relatively high probabilities of cluster 4 membership. This cluster consists entirely of write-offs. Thus rules 5 and 7 indicate high risk conditions. Rules 10, 13, 14, 16, 17, on the other hand, have relatively low probabilities of cluster 4, indicating low risk conditions. Note also that the high-risk rules also have fairly high membership in cluster 2 (profitable but often delinquent), whereas the low-risk rules have low membership in cluster 2.

Using these CART rules, new cardholder accounts are put into <u>one of 17 bins</u>, which has an impact on the optimal initial credit line for that account. The CART rules were validated using the holdout sample, by comparing the probabilities of cluster memberships for each rule's criteria in the validation set versus those in that rule.

Rule	Clus1	Clus2	Clus3	Clus4	Clus5
1	0.242	0.451	0.229	0.050	0.027
2	0.075	0.532	0.277	0.078	0.037
3	0.144	0.525	0.227	0.058	0.046
4	0.178	0.498	0.273	0.024	0.026
5	0.071	0.595	0.159	0.109	0.066
6	0.122	0.598	0.199	0.043	0.038
7	0.019	0.534	0.189	0.106	0.062
8	0.183	0.592	0.152	0.035	0.037
9	0.138	0.560	0.212	0.044	0.046
10	0.248	0.490	0.212	0.020	0.030
11	0.247	0.433	0.254	0.035	0.032
12	0.291	0.427	0.234	0.021	0.027
13	0.322	0.356	0.286	0.008	0.027
14	0.314	0.363	0.298	0.012	0.012
15	0.211	0.532	0.190	0.028	0.038
16	0.259	0.484	0.198	0.018	0.040
17	0.349	0.441	0.181	0.012	0.017

Table 1: Cluster membership probabilities for the 17 CART rules.

Initial Line Optimization

For each new credit card applicant, we calculated the expected NPV at each initial credit line CL as follows:

Exp(NPV | CL) =
  Pr(Cluster 1)* Exp(NPV | Cluster 1, CL)
+ Pr(Cluster 2)* Exp(NPV | Cluster 2, CL)
+ Pr(Cluster 3)* Exp(NPV | Cluster 3, CL)
+ Pr(Cluster 4)* Exp(NPV | Cluster 4, CL)
+ Pr(Cluster 5)* Exp(NPV | Cluster 5, CL)

Using this weighted average formula, we calculated the expected NPV at a variety of initial credit lines. The optimal credit line for a given rule is that with the maximum NPV; a unique optimum existed for each rule, with no distinct local optima. We don't show the detailed lines here; the highest lines were 67% higher than the lowest lines. The accounts in the higher risk rules (5 and 7) were assigned lower lines, whereas those in lower risk rules are assigned higher lines.

Conclusions and Future Work

We have presented a framework for credit customer optimization based on clustering and prediction. This framework is flexible in allowing various schemes. Other segmentation methods are possible, as well as other prediction techniques (such as neural networks). Below we describe other ways the basic framework may be extended.

Extending the Approach to Managing Credit Lines Over Time

The assignment of initial credit lines does not solve the entire business problem. There is still a need to determine the size and timing of credit line changes as customer behavior is observed. The existing framework can be extended to handle the dynamic problem by using Bayes' Rule to update the membership probabilities. Specifically, equation 1 below can be used to update the probability of being in cluster c given observed behavior $\underline{x}$, where $\underline{x}$ is a vector of discrete performance measures (say, spending patterns and payment behavior):

(1) $P(c \mid \underline{x}) = P(\underline{x} \mid c)*P(c)/P(\underline{x})$

Here, $P(\underline{x})$ and $P(\underline{x} \mid c)$ are estimated from the historical data. A Bayesian approach requires that the multivariate distributions $P(\underline{x})$ and $P(\underline{x} \mid c)$ be specified. Choosing a suitable family of multivariate distributions in this case is difficult for a number of reasons. First, the performance measures comprising $\underline{x}$ are not independent, nor are they of like type. Some measures may be integer (i.e. number of months delinquent), while others are continuous. Additionally, there is reason to believe that the distribution of $\underline{x}$ is time dependent. This is clearly the case when delinquency is a part of $\underline{x}$. Lastly, visual inspection of some spending measures reveal highly non-normal distributions.

For these reasons, we propose calculating the empirical distribution of $\underline{x}$ for various account ages t (in months), say t = 4, 8, 12, ... , 32. We propose discretizing $\underline{x}$ by binning the constituent performance measures (i.e. spending and payment measures) at appropriate levels. If $\underline{x}$ comprises three performance measures with four levels each, then there would be $4^3 = 64$ discrete values of $\underline{x}$. The more levels of each measure, the more data are required to accurately estimate $P(\underline{x})$ and $P(\underline{x} \mid c)$. In fact, this approach is subject to the explosion of dimensionality present in many techniques involving the discretization of continuous variables. The main problem is the amount of data to estimate $P(\underline{x} \mid c)$. However, the above scenario of 64 discrete values of $\underline{x}$ is reasonable given the large size of many customer databases.

Given $P(\underline{x})$ and $P(\underline{x} \mid c)$ at time t, equation (1) can be used to obtain an updated estimate of the probability of cluster membership for all clusters. In the above paragraph, we suppose that these probabilities would be updated every four months until an account is 32 months old. The updated membership probabilities would used as before to determine the new optimal credit line. If the new optimal credit line differs from the current line, then the appropriate line change would be recommended.

Extension to higher dimensions

We have demonstrated credit account optimization as a function of one independent variable, the initial credit line. However, one may wish to optimize credit line as a function of multiple independent variables, such as repayment terms, interest rates, etc.

In principle, our same clustering and prediction framework applies, but the challenges are in finding the optima of a multidimensional input space. The optima is unlikely to be unique, and there may not be sufficient data to accurately represent the entire space. These limitations may be overcome using search and optimization techniques in data-rich domains.

References

Bitran, G.R., and Mondschein, S.V., "Mailing Decisions in the Catalog Sales Industry," Management Science, v.42, no.9, September 1996, pp.1364-1381.

Brealy, R.A., and Myers, S.C., Principles of Corporate Finance, fourth edition, Mc-Graw Hill, 1991.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C. J., Classification and Regression Trees, Chapman & Hall, 1993.

Hughes, A. M., "Lifetime Value, the Criterion of Strategy," chapter 3 of Strategic Database Marketing, Irwin Professional Publishing, 1994.

Rosenberg, E. and Gleit, A., "Quantitative Methods in Credit Management: A Survey," Operations Research, v.42, n. 4, July-August 1994, pp. 589-613.

Acknowledgments

Bill Hunt, Michael Koukounas, Brian Murren, and Junjie Xiong of General Electric have all provided valuable effort on this project and this paper. Margaret Trench of General Electric has been especially supportive.