Bayesian Behavior Scoring Model

Ling-Jing Kao[1*], Fengyi Lin[1] and Chun Yuan Yu[2]
[1]*National Taipei University of Technology*
and [2]*National Chiao Tung University*

Abstract: Although many scoring models have been developed in literature to offer financial institutions guidance in credit granting decision, the purpose of most scoring models are to improve their discrimination ability, not their explanatory ability. Therefore, the conventional scoring models can only provide limited information in the relationship among customer demographics, default risk, and credit card attributes, such as APR (annual percentage rate) and credit limits. In this paper, a Bayesian behavior scoring model is proposed to help financial institutions identify factors which truly reflect customer value and can affect default risk. To illustrate the proposed model, we applied it to the credit cardholder database provided by one major bank in Taiwan. The empirical results show that increasing APR will raise the default probability greatly. Single cardholders are less accountable for credit card repayment. High income, female, or cardholders with higher education are more likely to have good repayment ability.

Key words: Bayesian approach, behavior scoring model, credit card.

1. Introduction

Financial institutions, such as credit card issuing banks, make heavy use of various scoring models to make a decision on a credit application and to monitor a consumer' repayment behavior. The model for the former purpose is known as a credit scoring model while the model for the later purpose is known as a behavior scoring model. Most scoring models are built for yielding a binary outcome indicating whether an application should be approved given the first-time applicant's debt ratio, credit report, and demographic variables. Regardless of purposes, the most common scoring models include logit model (e.g., Hand *et al.*, 2005), linear discrimination analysis (e.g., Lee *et al.*, 2002), data envelopment analysis (e.g., Min *et al.*, 2008), and other data mining approaches such as neural networks (e.g., West, 2000; Lee *et al.*, 2002; Baesens *et al.*, 2003; Baesens *et al.*,

---

[*]Corresponding author.

2003), classification trees (e.g., Paass and Kindermann, 1998), support vector machine approach (e.g., Baesens *et al.*, 2003; Huang *et al.*, 2007), and rule-based approaches (e.g., Messier and Hansen, 1988).

Even though these scoring models offer financial institutions guidance in credit granting decision, the current scoring models suffer from the following limitations. First, the scoring model can only reflect an applicant's current status. But, in reality, the repayment ability of a customer is varied with personal circumstances which cannot be observed by credit card issuing banks and researchers. The most financial institutions can do is to monitor a customer's repayment history over time. As long as the defaulting signal appears, the issuing bank can prepare for the consequence.

Second, the scoring model is constructed under the assumption that the credit-granting decision made by the financial institution is correct. But, how the financial institution makes the credit-granting decision is usually unknown and questionable. Third, most scoring model concerns the discrimination ability, not its explanatory ability. The discrimination ability of a scoring model is usually evaluated by type 1 and type 2 errors. However, as mentioned above, using the discrimination ability to evaluate the performance of scoring model is questionable because it is given that "true" decisions made by the financial institution is correct. In addition, because the model is lack of explanatory ability, it provides limited information on customer types and its relationship to product attributes, such as APR (annual percentage rate) and credit limits. Therefore, these scoring models with decent predictive accuracy still cannot help financial institutions determine factors which truly reflect customer value, and how credit line and APR should be set to minimize defaulting risk.

To address these issues, we present a Bayesian behavior scoring model to parameterize the relationship between customer repayment ability, credit card product attributes (e.g., credit limits and annual percentage rate (APR)), and demographic variable. To illustrate the proposed model, we applied it to the credit cardholder database provided by one major bank in Taiwan. This dataset includes the transaction records, demographic variables, and payment information of 2,948 cardholders from January 1, 2010 to December 31, 2010. We successfully demonstrate that the proposed Bayesian behavior scoring method can be employed to evaluate customer repayment ability given credit card attributes and demographic variables. Our methodology gives financial institutions additional option in analyzing customer value and monitoring customers who may have higher default risk.

This paper is organized as follows. Section 2 describes the proposed Bayesian behavior scoring model. Section 3 presents the empirical results. Concluding remarks is given in Section 4.

2 Bayesian Behavior Scoring Model

In this paper, we proposed a Bayesian behavior scoring model to study customer value. The rationality behind our model development is that consumer credit is granted given the likelihood that the borrower can repay the loan on time in the future. Thus, debt ratio should be an important factor for determining a credit applicant's repayment ability.

Let H denotes the total number of customers, w_{ij} denotes the ith customer's amount of principal repayment at the jth period; p_{ij} denotes the ith customer's total principal amount payable at the jth period; Q_{ij} denotes the ith customer's amount of interest paid at the jth period; s_{ij} denotes the ith customer's total interest payment due at the jth period. Then, the ith cardholder's repayment ability at the jth period is defined as

$$M_{ij} = \rho \left(\frac{w_{ij}}{p_{ij}} \right) + (1 - \rho) \left(\frac{Q_{ij}}{s_{ij}} \right). \tag{1}$$

In Eq.(1), the first term (w_{ij}/p_{ij}) represents a customer's principal repayment ability, and the second term (Q_{ij}/s_{ij}) is used to measure a customer's interest repayment. ρ $(0 \leq \rho \leq 1)$ represents the weight given to the importance of each repayment ability. Since M_{ij} is a weight average of two ratios, the value of M_{ij} is between 0 and 1. The greater M_{ij} is, the higher repayment ability the ith customer has.

To further determine how customers' credit limits (x_{1ij}) and APRs (x_{2ij}) affect customer repayment ability and explain the heterogeneity of customer repayment ability, we let x_{1ij} represent the ith credit cardholder's credit limit at the jth period, x_{2ij} represent the ith credit card holder's APR at the jth period, z_i represent the ith credit card holder's demographic variables. $M_{ij}/(1 - M_{ij} + k)$ is the odds of the customer repayment, where k is a small positive constant added to avoid having a zero denominator which occurs when a credit cardholder pays in full on time. The advantage of using the odds of the customer repayment is that we can measure the relative likelihood of being a valuable credit cardholder. For example, when the odds equals to 4, the likelihood of being an accountable customer is four times of the likelihood of being delinquent in payment. We let the behavior scoring model for the ith cardholder take the loglinear form and define it as:

$$\ln \left(\frac{M_{ij}}{1 - M_{ij} + k} \right) = \beta_{i0} + \beta_{i1} x_{1ij} + \beta_{i2} x_{2ij} + \varepsilon_{ij}, \tag{2}$$
$$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \ \ \forall j = 1, 2, \cdots, J_i,$$

where J_i denotes the total number of observation of the ith cardholder. Therefore,

the behavior scoring model defined in Eq.(2) can be expressed as

$$
\begin{pmatrix}
\ln\left(\frac{M_{i1}}{1-M_{i1}+k}\right) \\
\ln\left(\frac{M_{i2}}{1-M_{i2}+k}\right) \\
\vdots \\
\ln\left(\frac{M_{iJ_i}}{1-M_{iJ_i}+k}\right)
\end{pmatrix}
=
\begin{pmatrix}
1 & x_{1i1} & x_{2i1} \\
1 & x_{1i2} & x_{2i2} \\
\vdots & \vdots & \vdots \\
1 & x_{1iJ_i} & x_{2iJ_i}
\end{pmatrix}
\times
\begin{pmatrix}
\beta_{i0} \\
\beta_{i1} \\
\beta_{i2}
\end{pmatrix}
+
\begin{pmatrix}
\varepsilon_{i1} \\
\varepsilon_{i2} \\
\vdots \\
\varepsilon_{iJ_i}
\end{pmatrix},
\qquad (3)
$$

where $B_i = (\beta_{i0}, \beta_{i1}, \beta_{i2})'$ represents intercept and coefficients of product attributes credit limit (x_{1ij}) and APR(x_{2ij}). We further let

$$
Y_i = \left(\ln\left(\frac{M_{i1}}{1-M_{i1}+k}\right), \ln\left(\frac{M_{i2}}{1-M_{i2}+k}\right), \cdots, \ln\left(\frac{M_{iJ_i}}{1-M_{iJ_i}+k}\right) \right)',
$$

then Eq.(3) can be re-defined as

$$
\begin{aligned}
Y_i &=
\begin{pmatrix}
1 & x_{1i1} & x_{2i1} \\
1 & x_{1i2} & x_{2i2} \\
\vdots & \vdots & \vdots \\
1 & x_{1iJ_i} & x_{2iJ_i}
\end{pmatrix}
\times
\begin{pmatrix}
\beta_{i0} \\
\beta_{i1} \\
\beta_{i2}
\end{pmatrix}
+
\begin{pmatrix}
\varepsilon_{i1} \\
\varepsilon_{i2} \\
\vdots \\
\varepsilon_{iJ_i}
\end{pmatrix} \\
&= X_i B_i + e_i; \quad e_i \sim \mathcal{N}(0, \sigma^2 I).
\end{aligned}
\qquad (4)
$$

To further investigate the influence of demographic variables on customer response to credit card attributes, we further assume that

$$
B_i = \Theta z_i + \Delta_i, \quad \Delta_i \sim \mathcal{N}(0, \Lambda), \qquad (5)
$$

where z_i is a $d \times 1$ vector of the ith cardholder's demographic variables, Θ is a $3 \times d$ coefficient matrix, and Δ_i is a 3×1 vector of the ith cardholder's stochastic terms distributed as $\mathcal{N}_3(0, \Lambda)$.

The Gibbs sampler was used to estimate the proposed model given the following conjugate priors:

$$
\begin{aligned}
\text{vec}(\Theta) &\sim \mathcal{N}(u_0, V_0), \\
\Lambda &\sim \mathcal{IW}(f_0, G_0), \\
\sigma^2 &\sim \mathcal{IG}(r_0/2, s_0/2),
\end{aligned}
\qquad (6)
$$

where $\text{vec}(\Theta)$ stacks the columns of Θ which is a $3 \times d$ coefficient matrix. The detail explanation regarding the Gibbs sampler and conjugate priors can be found at Gelfand and Smith (1990) and Smith and Roberts (1993).

Markov Chain Monte Carlo (MCMC) procedure is carried out by generating draws from the following full conditional distributions:

$$[B_i \mid \text{rest}] \propto [Y_i \mid B_i, \sigma^2 I][B_i \mid \Theta, \Lambda]$$

$$\sim \mathcal{N}_3 \left(\left(\frac{1}{\sigma^2} X_i'X_i + \Lambda^{-1} \right)^{-1} \left(\frac{1}{\sigma^2} X_i'Y_i + \Lambda^{-1}\Theta'z_i \right), \left(\frac{1}{\sigma^2} X_i'X_i + \Lambda^{-1} \right)^{-1} \right), (7)$$

$$[\sigma^2 \mid \text{rest}] \propto \prod_{i=1}^{H} \left[Y_i \mid B_i, \sigma^2 \right] \left[\sigma^2 \mid r_0/2, s_0/2 \right]$$

$$\sim \mathcal{IG} \left(\frac{r_0 + H}{2}, \frac{s_0 + \sum_{i=1}^{H}(Y_i - X_iB_i)'(Y_i - X_iB_i)}{2} \right), \tag{8}$$

$$\text{Vec}(\Theta) \sim \mathcal{N} \left(W \cdot \left[\left(Z' \otimes \Lambda^{-1} \right) \text{Vec}(B') + V_0^{-1}u_0 \right], W \right),$$

where

$$W = \left[Z'Z \otimes \Lambda^{-1} + V_0^{-1} \right]^{-1},$$
$$Z = \left[z_1', z_2', \cdots, z_H' \right]'$$
$$= \begin{bmatrix} z_1' \\ z_2' \\ \vdots \\ z_H' \end{bmatrix}$$
$$= \begin{bmatrix} z_{10} & z_{11} & \cdots & z_{17} \\ z_{20} & z_{21} & \cdots & z_{27} \\ \vdots & \vdots & \ddots & \vdots \\ z_{H0} & z_{H1} & \cdots & z_{H7} \end{bmatrix}, \tag{9}$$

$$[\Lambda \mid \text{rest}] \propto \prod_{i=1}^{H} [B_i \mid \Theta, \Lambda] [\Lambda \mid f_0, G_0]$$

$$\sim \mathcal{IW} \left(H + f_0, \left(\sum_{i=1}^{H}(B_i - \Theta z_i)(B_i - \Theta z_i)' + G_0 \right) \right). \tag{10}$$

The calculation details of the full conditionals above are given in the appendix.

3 Empirical Study

A dataset provided by a leading bank in Taipei, Taiwan, is used to illustrate the proposed Bayesian behavior scoring model. This dataset contains the information of credit cardholders whose applications were approved between 2008 and

2009 and have transaction records spanned over a complete year from January 1, 2010 to December 31, 2010. The transaction data contains account-level information, such as the transaction date, the amount of transaction, repayment date and the amount of repayment.

To ensure the consistency of data field and the requirement of Bayesian behavior scoring model, data preprocessing was needed. We first used a customer identification to match two raw datasets: one contains effective credit card account information of 25,328 credit cardholders, and the other stores over 5 million individual transaction records for these accounts. To satisfy the requirement of Bayesian model estimation, we only kept those credit card holders who have at least 8 repayment records within one-year period. In this case, 8,915 credit card holders were obtained. Then, we randomly selected 2,948 credit card holders, which represent 25% of the valid sample, to build the behavioral scoring model. In this sample of 2,948 credit card holders, who have at least 8 repayment records within a one-year period, each cardholder has his/her five demographic variables, annual income, and credit limits with APR granted by the bank. The summary statistics for variables are presented in Table 1.

The model was evaluated by assuming $\rho = 0.9$, $\rho = 0.5$, and $\rho = 0.1$ for different decision scenarios. When $\rho = 0.5$, equal weights were assigned to a customer's principal repayment ability and a customer's interest repayment ability. When $\rho = 0.9$, a customer's principal repayment ability is the major concern of the credit card issuer. On the other hand, when $\rho = 0.1$, a customer's interest repayment ability is much more important than a customer's principal repayment ability. For all scenarios, the MCMC ran for a total of 2,000 iterations and converges after 1,500 MCMC iterations. The draws of last 500 iterations were used to compute posterior estimates.

3.1 The Posterior Estimates of B_i

To better demonstrate the result, we computed the average of the posterior draw of B_i by letting $\overline{\beta} = \sum_{i=1}^{H} B_i / H$. The posterior estimates $\overline{\beta}$ and $\exp(\overline{\beta})$ are presented in Tables 2 and 3, respectively. The purpose of taking exponential on $\overline{\beta}$ is to exam the impact of its corresponding product attribute on the likelihood of being default and the likelihood of being responsible.

As shown in Table 2, for all scenarios, all estimated value of $\overline{\beta}$ are away from zero, indicating that both credit limit and APR have significant impacts on the odds of customer repayment. Table 3 shows that, when APR and credit limits are not considered, the average likelihood of being a responsible customer is 1.68, 10.09 times, and 19.74 times of being a default customer for $\rho = 0.1$, $\rho = 0.5$, and $\rho = 0.9$ respectively.

Table 1: Summary statistics for the variables

Gender	Annual Percentage Rate (APR)
1 = female (51.7%)	Mean = 17%
2 = male (49.3%)	Std deviation = 0.041

Marriage Status	Credit Limits
1 = single (42.5%)	Mean = 16.1385
2 = married (52%)	Std deviation = 107124

Annual income
  Mean = 490108.42
  Std deviation = 298625.38

Education level
  1. Ph.D. (0.2%)
  2. Post-graduate (2%)
  3. College (8.9%)
  4. Vocational school (21.6%)
  5. High school (44.1%)
  6. Other (23.1%)

Industry sector
  1. Farming, forestry and fisheries
  2. Manufacturing
  3. Construction
  4. Transportation and Communications
  5. Business service
  6. Public administration
  7. Mining and stone
  8. Wholesale trade
  9. Finance and insurance
  10. Social service and personal service
  11. Others

Work class
  1. White-collar
  2. Professional
  3. Solider, public school teacher, and government officers
  4. Blue-collar
  5. Service and others

Table 2: $\overline{\beta}$ posterior estimates

	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.9$
$\overline{\beta}_0$ (Intercept)	0.519 (0.692)	2.312 (0.727)	2.983 (0.634)
$\overline{\beta}_1$ (credit limit)	0.330 (0.185)	0.513 (0.166)	0.900 (0.240)
$\overline{\beta}_2$ (APR)	-3.855 (1.056)	-4.918 (1.268)	-5.675 (1.109)

Table 3: $\exp(\overline{\beta})$ posterior estimates

	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.9$
$\exp(\overline{\beta}_0)$ (Intercept)	1.68	10.09	19.75
$\exp(\overline{\beta}_1)$ (credit limit)	1.39	1.67	2.46
$\exp(\overline{\beta}_2)$ (APR)	0.02	0.007	0.003

When one unit of credit limit is granted, the average likelihood of being a responsible customer is 1.39, 1.67, and 2.46 times of being a default customer given everything else remains the same for $\rho = 0.1$, $\rho = 0.5$, and $\rho = 0.9$ respectively. Finally, when APR is increased by one unit, the average of likelihood of being a responsible customer is 0.02, 0.007, and 0.003 times of being a default customer given everything else remains the same.

Regardless of the scenarios, the following pattern can be observed in Table 3. When the value of ρ increases, the values of $\exp(\overline{\beta}_0)$ and $\exp(\overline{\beta}_1)$ increase, but the value of $\exp(\overline{\beta}_2)$ decreases. This result indicates that increasing APR will raise the default probability greatly because the high incurring interest will make a cardholder even more difficult to reduce his/her total debt. The positive association between the likelihood of being accountable and credit limit is due to the fact that higher credit limit is usually assigned to an applicant with better credit score and higher income. Thus, this result is consistent to the general expectation that an applicant with higher income can usually keep up regular payment. Finally, the change in $\exp(\overline{\beta}_0)$ for different values of ρ shows the critical impact of a credit card issuer's policy on the likelihood of being accountable. A more conservative policy is implemented if a credit card issuer's concern a credit cardholder's principal repayment ability more (e.g., $\rho = 0.9$) and appreciate those who can pay principal regularly. If a credit card holder can pay most of principal, less interest will incur, and he has less chance to be default.

3.2 Posterior Estimates of Θ

The influence of each demographic variables (Z_d) on the coefficients (β_{i0}, β_{i1}, β_{i2}) of credit card attributes for each scenario is summarized in Table 4. As shown in Table 4, if the weight assigned to the principal repayment ability increase (e.g., $\rho = 0.9$), demographic variables have less explanatory ability in the coefficients (β_{i0}, β_{i1}, β_{i2}) of credit card attributes. For example, for the scenario of $\rho = 0.9$ in Table 4(c), only gender has influence on the coefficient of APR (β_{i2}). However, for the scenario of $\rho = 0.1$ in Table 4(a), annual income, gender, industry sector, and education have significant influence on β_{i0}, while the coefficient of credit limit (β_{i1}) is influenced by intercept, work class, and industry type, and the coefficient of APR (β_{i2}) is influenced by gender and education.

Table 4: Posterior estimates of Θ

(a) $\rho = 0.1$

Θ	Z_0 Intercept	Z_1 Age	Z_2 Annual Income	Z_3 Gender	Z_4 Work class	Z_5 Industry	Z_6 Marital Status	Z_7 Education
β_{i0}	-4.2912 (1.5159)	0.0189 (0.0196)	**-0.1421** **(0.0208)**	**1.0911** **(0.2294)**	-0.0024 (0.0004)	**0.0011** **(0.0002)**	-0.4714 (0.1051)	**0.7921** **(0.0950)**
β_{i1}	**0.5564** **(0.1747)**	0.0022 (0.0046)	0.0298 (0.0104)	-0.1170 (0.0770)	**0.0006** **(0.0001)**	**-0.0002** **(0.0000)**	-0.0433 (0.0712)	-0.0552 (0.0283)
β_{i2}	1.4779 (0.6720)	-0.0365 (0.0131)	0.0269 (0.0286)	**-0.8206** **(0.2490)**	0.0005 (0.0004)	-0.0008 (0.0003)	0.3337 (0.1725)	**-0.6274** **(0.0914)**

(b) $\rho = 0.5$

Θ	Z_0 Intercept	Z_1 Age	Z_2 Annual Income	Z_3 Gender	Z_4 Work class	Z_5 Industry	Z_6 Marital Status	Z_7 Education
β_{i0}	0.0602 (1.2069)	-0.02882 (0.0284)	**-0.14607** **(0.0486)**	0.84725 (0.3315)	-0.00124 (0.0010)	0.00082 (0.0005)	0.40911 (0.2928)	0.39431 (0.1578)
β_{i1}	0.23192 (0.2022)	0.00465 (0.0066)	0.02718 (0.0081)	-0.00023 (0.0819)	0.00045 (0.0003)	-0.00004 (0.0001)	**-0.1319** **(0.0327)**	0.01984 (0.0347)
β_{i2}	-0.2941 (0.7996)	-0.00692 (0.0149)	0.02532 (0.0293)	**-0.84454** **(0.1790)**	-0.00004 (0.0005)	-0.00067 (0.0004)	-0.0687 (0.2120)	**-0.54971** **(0.1001)**

(c) $\rho = 0.9$

Θ	Z_0 Intercept	Z_1 Age	Z_2 Annual Income	Z_3 Gender	Z_4 Work class	Z_5 Industry	Z_6 Marital Status	Z_7 Education
β_{i0}	1.5266 (1.4455)	-0.0567 (0.0253)	-0.1139 (0.1078)	0.8308 (0.4055)	-0.0017 (0.0012)	0.0016 (0.0006)	0.4912 (0.4143)	0.2991 (0.1994)
β_{i1}	0.4541 (0.2641)	0.0073 (0.0042)	0.0095 (0.0231)	0.1163 (0.1169)	0.0003 (0.0002)	-0.0002 (0.0001)	-0.1575 (0.0699)	0.0645 (0.0434)
β_{i2}	-0.7995 (0.7446)	0.0063 (0.0140)	0.0160 (0.0479)	**-1.0051** **(0.1889)**	0.0002 (0.0006)	-0.0008 (0.0003)	-0.0606 (0.2315)	-0.6395 (0.1019)

Table 4 also shows that gender is the only demographic variable which is significant regardless of the value of ρ. The negative value $\Theta_{3,4}$ (e.g., $\Theta_{3,4} = -1.0051$ for $\rho = 0.9$) shows that, compared to the male cardholder ($Z_3 = 2$), the

female cardholder ($Z_3 = 1$) are more likely to have good repayment ability for all scenarios. For the scenario of $\rho = 0.5$ and $\rho = 0.1$, annual income (Z_2) is the only variable which influence the coefficient of intercept β_{i0}, and gender (Z_3) and education (Z_7) are only variables which influences the coefficient of APR (β_{i2}). The negative value of $\Theta_{3,4}$ (gender) and $\Theta_{3,8}$ (education) indicates that female credit cardholders with higher education are less likely to be default given the same APR is assigned.

For $\rho = 0.1$, the male customer who has less education, less income, and work at the service area is more likely to keep up with the repayment given the fixed APR and credit limits. This result is reasonable for the scenario of $\rho = 0.1$ because those customers may not have ability to pay most of the principal, but can pay interest regularly on time. In addition, work class (Z_4) and industry sector (Z_5) have little influence on the coefficient of credit limit (β_{i1}) since, in Table 4(a), the posterior estimates of $\Theta_{2,5}$ and $\Theta_{2,6}$ are very close to zero.

For $\rho = 0.5$, except annual income (Z_2) and marital status (Z_6), demographic variables have no significant effect on credit limit. This result is reasonable because credit limit is usually granted according to annual income. Besides, this result also suggests that high income cardholders tends to have a greater coefficient of credit limit (β_{i1}), which implies that high income cardholders have higher likelihood to be accountable for both principal and interest repayment. Compared to the married cardholders, the unmarried cardholders are more sensitive to their credit limit, and they are less accountable for credit card repayment.

3.3 Posterior Estimates of Λ

Table 5 shows the posterior estimates of Λ for $\rho = 0.1$, $\rho = 0.5$, and $\rho = 0.9$ respectively. The covariance estimates are given in the upper triangular of the matrix, and the correlation estimates are given in the lower triangular matrix. The standard deviation of the estimates is shown in parentheses.

Table 5: Posterior estimates of Λ

(a) $\rho = 0.1$

Λ	Intercept	Credit Limit	APR
Intercept	2.2831 (1.1657)	-0.2753 (0.1317)	-4.6373 (1.2880)
Credit Limit	-0.6450	0.0798 (0.0061)	-0.6095 (0.0792)
APR	-0.9599	-0.669	10.4006 (0.4717)

Table 5 (continued): Posterior estimates of Λ

(b) $\rho = 0.5$

Λ	Intercept	Credit Limit	APR
Intercept	6.62067 (1.8646)	-0.66135 (0.2616)	-5.11775 (1.17704)
Credit Limit	-0.4753	0.2924 (0.02819)	-0.596 (0.12831)
APR	-0.6253	-0.3465	10.11703 (0.63669)

(c) $\rho = 0.9$

Λ	Intercept	Credit Limit	APR
Intercept	22.0563 (2.0858)	-3.6656 (1.4567)	-9.0902 (2.9487)
Credit Limit	-0.9285	0.7067 (0.0887)	-0.9612 (0.2244)
APR	-0.6123	-0.3617	9.9140 (0.6197)

As shown in Table 5, the variance of intercept increases when the value of ρ increases. It suggests that, when the bank places more weight on the principal repayment ability, the estimates of β_{i0} become more heterogeneous and cardholders have various levels of repayment abilities. So, placing more weight on the principal repayment ability becomes a better way to distinguish customer types. Even though the negative correlation estimates between intercept and APR at different values of ρ can be observed, the scenario of $\rho = 0.9$, relative to those of $\rho = 0.1$ and $\rho = 0.5$, has stronger negative correlation. The negative correlation between intercept and APR is also observed. The stronger negative correlation is observed when $\rho = 0.1$, and the estimated values for the scenarios of $\rho = 0.5$ and $\rho = 0.9$ are close to each other.

Even though no clear pattern can be observed from the variance of APR and the covariance between APR and Credit Limit, we view the variance of APR in all three scenarios indifferent because all estimated values are close to 10. For the covariance between APR and Credit Limit, the estimated values for the scenarios of $\rho = 0.1$ and $\rho = 0.5$ are close to -0.6 while the estimated values $\rho = 0.9$ becomes -0.9612. This pattern described above suggests that greater consumer heterogeneity is observed when more weight is placed on the principal repayment ability ($\rho = 0.9$). This result also suggests that the principal repayment ability can be a better way to discriminate customer's repayment ability.

The negative association is observed between the coefficient of APR (β_2) and the coefficient of intercept (β_0), and the coefficient of APR (β_2) and the coefficient of credit limits (β_1) in all scenarios. The negative correlation between the coefficient of APR (β_2) and the coefficient of credit limits (β_1) is supported by Table 2 in which the posterior mean $\overline{\beta}_1$ and $\overline{\beta}_2$ have opposite sign. The same explanation can be applied to the negative association between the coefficient of intercept (β_0) and the coefficient of APR (β_2) as well. The negative association between the coefficient of intercept (β_0) and he coefficient of credit limit (β_1) is observed even though both $\overline{\beta}_0$ and $\overline{\beta}_1$ have positive values as indicated in Table 2. We believe that this occur because of the greater variance of intercept (β_0), which leads to a greater variance in each individual's intercept coefficient estimates.

3.4 Summary

In this empirical study, we have demonstrated that the proposed Bayesian behavior scoring method can be employed to evaluate the odds of customer repayment ability given credit card attributes. Moreover, the proposed model can successfully identify factors which will lead to higher default probability. Therefore, financial institutions can have additional option in analyzing customer value and monitoring customers who may have higher defaulting risk.

4 Concluding Remarks

Even though inappropriate credit granting decisions can result in high credit risk, monitoring and controlling the default risk after credit is granted is equally important. In this paper, a Bayesian scoring model is constructed to parameterize relationship among the odds of being accountable, financial product attributes, and customer demographics. Unlike most scoring models developed in literature which either focuses on improving the discrimination ability and prediction accuracy of a model or deriving the decision rule for credit granting, our methodology can help the financial institutions identify factors which can lead to higher default risk. As long as the defaulting signal appears, the credit issuers can take action and prepare for the consequence.

A credit card dataset provided by a local bank in Taiwan is used to illustrate the proposed Bayesian behavior scoring model under three different scenarios. In general the empirical result shows that both credit limit and APR have significant impacts on the odds of customer repayment. Specially increasing APR will raise the default probability greatly. This result is reasonable because the high incurring interest will make a cardholder even more difficult to reduce his total debt. The unmarried cardholders are more sensitive to their credit limit, and they are less accountable for credit card repayment. High income customers have higher

likelihood to be accountable for credit card repayment. Female cardholders and cardholders with higher education are more likely to have good repayment ability.

This research demonstrates a practical Bayesian application in monitoring customer default risk. The model itself is straightforward and can be extended easily to other financial products, such as mortgage or car loan. In future research, the objective function, such as profit maximization or risk minimization, can be incorporated into the proposed model to design credit granting policies.

Acknowledgements

Appendix

Cardholder: $i = 1, 2, \cdots, H$.
Observations of cardholder i: $j = 1, 2, \cdots, J_i$.
Y_i: a J_i by 1 vector of the ith cardholder's repayment ability.
X_i: a J_i by M design matrix with intercept. In the empirical study, M equals to 3.
B_i: an M by 1 vector of coefficients corresponding to X_i.
e_i: a J_i by 1 vector of the ith cardholder's residuals.
z_i: a d by 1 demographic vector of the ith cardholder's and

$$
Z = \left[z_1', z_2', \cdots, z_H' \right]' = \begin{bmatrix} z_1' \\ z_2' \\ \vdots \\ z_H' \end{bmatrix} = \begin{bmatrix} z_{10} & z_{11} & \cdots & z_{17} \\ z_{20} & z_{21} & \cdots & z_{27} \\ \vdots & \vdots & \ddots & \vdots \\ z_{H0} & z_{H1} & \cdots & z_{H7} \end{bmatrix}.
$$

Model:

$$
Y_i = \left(\ln\left(\frac{M_{i1}}{1 - M_{i1} + k} \right), \ln\left(\frac{M_{i2}}{1 - M_{i2} + k} \right), \cdots, \ln\left(\frac{M_{iJ_i}}{1 - M_{iJ_i} + k} \right) \right)',
$$
$$
e_i = (\varepsilon_{i1}, \varepsilon_{i2}, \cdots, \varepsilon_{iJ_i})',
$$
$$
Y_i = X_i B_i + e_i, \quad e_i \sim \mathcal{N}(0, \sigma^2 I_{J_i}),
$$
$$
B_i = (\beta_{i,1}, \beta_{i,2}, \cdots, \beta_{i,M})' = \Theta z_i + \Delta_i; \quad \Delta_i \sim \mathcal{N}_M(0, \Lambda).
$$

Prior:

$$
\begin{aligned}
\mathrm{vec}(\Theta) &\sim \mathcal{N}(u_0, V_0), \\
\Lambda &\sim \mathcal{IW}(f_0, G_0), \\
\sigma^2 &\sim \mathcal{IG}(r_0/2, s_0/2).
\end{aligned}
$$

The full conditionals for MCMC can be obtained as follows:
(1) Draw B_i

$$
\begin{aligned}
[B_i \mid \mathrm{rest}] &\propto \left[Y_i \mid B_i, \sigma^2 I\right] \cdot [B_i \mid \Theta, \Lambda] \\
&\propto \exp\left\{-\frac{1}{2}\left(Y_i - X_i B_i\right)'\left(\sigma^2 I\right)^{-1}\left(Y_i - X_i B_i\right)\right\} \\
&\quad \cdot \exp\left\{-\frac{1}{2}\left(B_i - \Theta z_i\right)'\Lambda^{-1}\left(B_i - \Theta z_i\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\left(Y_i - X_i B_i\right)'\left(\sigma^2 I\right)^{-1}\left(Y_i - X_i B_i\right) + \left(B_i - \Theta z_i\right)'\Lambda^{-1}\left(B_i - \Theta z_i\right)\right]\right\} \\
&\sim \mathcal{N}_M\left(\left(\frac{1}{\sigma^2}X_i'X_i + \Lambda^{-1}\right)^{-1}\left(\frac{1}{\sigma^2}X_i'Y_i + \Lambda^{-1}\Theta z_i\right), \left(\frac{1}{\sigma^2}X_i'X_i + \Lambda^{-1}\right)^{-1}\right).
\end{aligned}
$$

(2) Draw σ^2

$$
\begin{aligned}
[\sigma^2 \mid \mathrm{rest}] &\propto \prod_{i=1}^{H}\left[Y_i \mid B_i, \sigma^2 I\right] \cdot \left[\sigma^2 \,\middle|\, \frac{r_0}{2}, \frac{s_0}{2}\right] \\
&\propto \prod_{i=1}^{H}\frac{1}{|\sigma^2 I|}\exp\left\{-\frac{1}{2}\left(Y_i - X_i B_i\right)'\left(\sigma^2 I\right)^{-1}\left(Y_i - X_i B_i\right)\right\} \\
&\quad \cdot \left(\sigma^2\right)^{-(r_0/2)-1}\exp\left(-\frac{s_0/2}{\sigma}\right) \\
&\propto \left(\sigma^2\right)^{-(H/2+r_0/2)-1}\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{H}\left(Y_i - X_i B_i\right)'\left(Y_i - X_i B_i\right) + s_0\right]\right\} \\
&\sim \mathcal{IG}\left(\frac{r_0 + H}{2}, \frac{s_0 + \sum_{i=1}^{H}\left(Y_i - X_i\beta_i\right)'\left(Y_i - X_i\beta_i\right)}{2}\right).
\end{aligned}
$$

(3) Draw Θ

Let

$$
\begin{bmatrix} B'_1 \\ B'_2 \\ \vdots \\ B'_H \end{bmatrix}_{H \times M}
=
\begin{bmatrix} \beta_{1,1} & \beta_{1,2} & \cdots & \beta_{1,M} \\ \beta_{2,1} & \beta_{2,2} & \cdots & \beta_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{H,1} & \beta_{H,2} & \cdots & \beta_{H,M} \end{bmatrix}
$$

$$
=
\begin{bmatrix} z'_1 \\ z'_2 \\ \vdots \\ z'_H \end{bmatrix}_{H \times d}
\cdot \Theta'_{d \times M}
+
\begin{bmatrix} \Delta'_1 \\ \Delta'_2 \\ \vdots \\ \Delta'_H \end{bmatrix}_{H \times M},
$$

$$
B = Z \cdot \Theta' + \Delta,
$$
$$
B' = \Theta \cdot Z' + \Delta'.
$$

Since

$$
\mathrm{Vec}(B') = \mathrm{Vec}\left(\Theta \cdot Z'\right) + \mathrm{Vec}(\Delta') = (Z \otimes I_M) \cdot \mathrm{Vec}(\Theta) + \mathrm{Vec}(\Delta'),
$$

and

$$
\mathrm{Vec}(B') \sim \mathcal{N}\left((Z \otimes I_M) \cdot \mathrm{Vec}(\Theta), I_H \otimes \Lambda\right).
$$

So, the full conditional of Θ is

$[\Theta \mid \text{rest}]$

$$
\propto \prod_{i=1}^{H} [B_i \mid \Theta z_i, \Lambda] \cdot \left[\mathrm{Vec}(\Theta) \mid u_0, V_0\right]
$$

$$
\propto \left[\mathrm{Vec}(B') \mid (Z \otimes I_M) \cdot \mathrm{Vec}(\Theta), I_H \otimes \Lambda\right] \cdot \left[\mathrm{Vec}(\Theta) \mid u_0, V_0\right]
$$

$$
\propto \exp\left\{-\frac{1}{2} \left[\mathrm{Vec}(B') - (Z \otimes I_M) \cdot \mathrm{Vec}(\Theta)\right]' (I_H \otimes \Lambda)^{-1}\right.
$$

$$
\left. \cdot \left[\mathrm{Vec}\left(B'\right) - (Z \otimes I_M) \cdot \mathrm{Vec}(\Theta)\right]\right\} \exp\left\{-\frac{1}{2} \left[\mathrm{Vec}(\Theta) - u_0\right]' V_0^{-1} \left[\mathrm{Vec}(\Theta) - u_0\right]\right\}
$$

$$
\cdot \left[\mathrm{Vec}(B') - (Z \otimes I_M) \cdot \mathrm{Vec}(\Theta)\right]' (I_H \otimes \Lambda)^{-1} \left[\mathrm{Vec}\left(B'\right) - (Z \otimes I_M) \cdot \mathrm{Vec}(\Theta)\right]
$$

$$
+ \left[\mathrm{Vec}(\Theta) - u_0\right]' V_0^{-1} \left[\mathrm{Vec}(\Theta) - u_0\right]
$$

$$
\propto \mathrm{Vec}(\Theta)' \left[(Z \otimes I_M)' (I_H \otimes \Lambda)^{-1} (Z \otimes I_M) + V_0^{-1}\right] \mathrm{Vec}(\Theta) - 2\mathrm{Vec}(\Theta)'
$$

$$
\cdot \left[(Z \otimes I_M)' (I_H \otimes \Lambda)^{-1} \mathrm{Vec}(B') + V_0^{-1} u_0\right]
$$

$$
\propto \mathrm{Vec}(\Theta)' \left[Z'Z \otimes \Lambda^{-1} + V_0^{-1}\right] \mathrm{Vec}(\Theta) - 2\mathrm{Vec}(\Theta)' \left[\left(Z' \otimes \Lambda^{-1}\right) \mathrm{Vec}\left(B'\right) + V_0^{-1} u_0\right].
$$

Thus, the posterior draw of Θ can be generated from

$$\text{Vec}(\Theta) \sim \mathcal{N}\left(W \cdot \left[\left(Z' \otimes \Lambda^{-1}\right)\text{Vec}(B') + V_0^{-1}u_0\right], W\right),$$

where $W = \left[Z'Z \otimes \Lambda^{-1} + V_0^{-1}\right]^{-1}$.

(4) Draw Λ

$$[\Lambda \mid \text{rest}] \propto \prod_{i=1}^{H}\left[B_i \mid \Theta z_i, \Lambda\right] \cdot \left[\Lambda \mid f_0, G_0\right]$$

$$\propto \prod_{i=1}^{H}\frac{1}{|\Lambda|^{1/2}}\exp\left\{-\frac{1}{2}\left(B_i - \Theta z_i\right)'\Lambda^{-1}\left(B_i - \Theta z_i\right)\right\} \cdot |\Lambda|^{-(f_0+M+1)/2}$$

$$\cdot \exp\left\{-\frac{1}{2}\text{tr}\left(G_0\Lambda^{-1}\right)\right\}$$

$$\propto \frac{1}{|\Lambda|^{H/2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{H}\left(B_i - \Theta z_i\right)'\Lambda^{-1}\left(B_i - \Theta z_i\right)\right\} \cdot |\Lambda|^{-(f_0+M+1)/2}$$

$$\cdot \exp\left\{-\frac{1}{2}\text{tr}\left(G_0\Lambda^{-1}\right)\right\}$$

$$= |\Lambda|^{-(H+f_0+M+1)/2}\exp\left\{-\frac{1}{2}\left(\text{tr}\Lambda^{-1}\left(\sum_{i=1}^{H}\left(B_i - \Theta z_i\right)\left(B_i - \Theta z_i\right)' + G_0\right)\right)\right\}$$

$$\sim \mathcal{IW}\left(H + f_0, \left(\sum_{i=1}^{H}\left(B_i - \Theta z_i\right)\left(B_i - \Theta z_i\right)' + G_0\right)\right).$$

References

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J. and Van-thienen, J. (2003). Bechmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society* **54**, 627-635.

Baesens, B., Setiono, R., Mues, C. and Vanthienen, J. (2003). Using neural network rule extraction and decision tables for credit-risk evaluation. *Management Science* **49**, 312-329.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398-409.

Smith, A. F. M. and Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the American Statistical Association* **55**, 3-23.

Hand, D. J. and Crowder, M. J. (2005). Measuring customer quality in retail banking. *Statistical Modelling* **5**, 145-158.

Huang, C. L., Chen, M. C. and Wang, C. J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications* **33**, 847-856.

Lee, T. S., Chiu, C. C., Chou, Y. C. and Lu, C. J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics & Data Analysis* **50**, 1113-1130.

Lee, T. S., Chiu, C. C., Lu, C. J. and Chen, I. F. (2002). Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications* **23**, 245-254.

Messier, W. F., Jr. and Hansen, J. V. (1988). Inducing rules for expert systems development: an example using default and bankruptcy data. *Management Science* **34**, 1403-1415.

Min, J. H. and Lee, Y. C. (2008). A practical approach to credit scoring. *Expert Systems with Applications* **35**, 1762-1770.

Paass, G. and Kindermann, J. (1998). Bayesian classification trees with overlapping leaves applied to credit-scoring. *Research and Development in Knowledge Discovery and Data Mining Lecture Notes in Computer Science* **1394**, 234-245.

West, D. (2000). Neural network credit scoring models. *Computers & Operations Research* **27**, 1131-1152.

Ling-Jing Kao
Department of Business Management
National Taipei University of Technology
No. 1, Sec. 3, Zhongxiao E. Rd., Taipei City 106, Taiwan
lingjingkao@ntut.edu.tw

Fengyi Lin
Institute of Commerce Automation and Management
National Taipei University of Technology
No. 1, Sec. 3, Zhongxiao E. Rd., Taipei City 106, Taiwan
fengyi@ntut.edu.tw

Chun Yuan Yu
Institute of Business and Management
National Chiao Tung University
No.118, Sec. 1, Jhongsiao W. Rd., Jhongjheng District, Taipei City 100, Taiwan
gogoafrica@gmail.com