Larger Datasets Lead to More Inaccurate Credit Scoring

Mimi Chong and Matt Davison
Department of Statistical and Actuarial Sciences,
Western University Canada, London, ON, CANADA
Corresponding author: Mimi Chong, e-mail: mchong9@stats.uwo.ca

Abstract

Credit scoring is an automated, objective and consistent tool which helps lenders to provide quick loan decisions. It can replace some of the more mechanical work done by experienced loan officers whose decisions are intuitive but potentially subject to bias. Existing credit scoring models are built using as many historical data as possible from past borrowers with known repayment performance. Analysts believe that using a larger dataset to build credit scoring models will always increase model accuracy. However, previous findings show that increasing the amount of data used to build models may result in a more complex model with no significant increase in accuracy. We will show that if some borrowers respond untruthfully to some questions, using higher dimensional data may even reduce the model's predictive power compared against using a dataset with lower dimensions. Using more data to build the model will increase the associated accumulated error and results in an overestimated model with low accuracy. The proposed issue will be studied using simulated data and discriminant analysis based on the credit scoring context. Knowing the optimal amount of data that is required to build credit scoring models can both improve accuracy and reduce operational cost and processing time. This research can help lending financial institutions to maximize profit in their loan activities.

Keywords: classification, discriminant analysis, lying, normal distribution

1 Introduction

As described in Anderson (2007), credit allows borrowers to buy now and pay later. Borrowers must show themselves trustworthy for instant by always repaying their loans according to the agreed terms, and providing sufficient information for the banks to grant credit decisions. Trust may also be enhanced by providing collateral. The amount of credit that will be granted varies among people. People with higher than average risk may still be granted credit in exchange for higher interest rates, lower credit limits and shorter repayment periods.

Credit scoring refers to the use of a numerical tool to rank borrowers according to their repaying desirability as a loan counterparty; it also ensures objective and consistent decisions. Characteristics of borrowers are collected and integrated into values which can be combined and interpreted through a credit scoring model. The model will then analyse all the information and provides an estimate of the default probability of borrowers. Credit scoring models present a shift from secured lending, based on collateral and guarantees, to unsecured lending, which relies on historical information and past experience.

It is reasonable to believe that using larger datasets will improve the predictive power of the resulting credit scoring models. Currently, all credit scoring models are built using as many historical data as possible. The purpose of this paper is to show that, if prospective borrowers lie in their responses to one or more questions, using higher dimensional data to build credit scoring models may lower the prediction power, compared against using lower dimensional

data. In fact, if a smaller dataset can be used to produce accurate predictions, the questionnaire used to collect information from borrowers can be shortened, which reduces the cost spend on collecting, checking and processing the data. This enhances the efficiency of the granting decisions.
The succeeding sections are organized as follows. Section 2 presents the method and model used for this problem. In section 3, the data used to analyze the problem was simulated. In section 4, we applied classification techniques to section 3's data to support the results. Conclusions are drawn in section 5.

2 Method and Model

This paper uses discriminant analysis, a well-known method in the area of credit scoring. According to Thomas, Edelman and Crook (2002) and Thomas (2000), the objective of discriminant analysis is to find a model that classifies all borrowers into two subsets, A_G and A_B, representing good and bad payers respectively. A model will be built using past payers with known repayment performance, and will be used to determine the granting decisions of new borrowers. A vector of variables $\mathbf{X} = (X_1, X_2, ..., X_p)$, representing the characteristics of each past payer, will be used to construct the model. We assume that p_G is the proportion of applicants who are good payers, and p_B the proportion of applicants who are bad payers. Let $p(\mathbf{x}|G)$ be the probability that a good borrower will have attribute $\mathbf{x}$, and $p(\mathbf{x}|B)$ be the probability that a bad borrower will have attribute $\mathbf{x}$. Two types of costs correspond to the two types of errors that can be made in this classification method. If a good payer was classified as a bad payer, the potential profit that the lender might have earned is lost. Conversely, if a bad payer was classified as a good payer, default losses will be incurred. We assume that the expected misclassification profit and loss incurred, denoted as L and D respectively, are the same for all borrowers.
The intention of the lender is to minimize the expected loss and maximize the expected profit. If a borrower was classified into A_G, there is only a loss if it is a bad payer, and the expected loss is $Dp(\mathbf{x}|B)p_B$. On the other hand, if a borrower was put into A_B, there is only a loss if it is a good payer, now with expected size $Lp(\mathbf{x}|G)p_G$. Therefore, we should classify a borrower into A_G if $Dp(\mathbf{x}|B)p_B \leq Lp(\mathbf{x}|G)p_G$. This condition leads to the set

$$A_G = \{\mathbf{x}|Dp(\mathbf{x}|B)p_B \leq Lp(\mathbf{x}|G)p_G\} \tag{1}$$

$$= \left\{\mathbf{x}|\frac{Dp_B}{Lp_G} \leq \frac{p(\mathbf{x}|G)}{p(\mathbf{x}|B)}\right\} \tag{2}$$

$$= \left\{\mathbf{x}|\frac{Dp_B}{Lp_G} \leq \frac{f(\mathbf{x}|G)}{f(\mathbf{x}|B)}\right\} \tag{3}$$

where the last equality follows if $\mathbf{x}$ is a vector of continuous characteristic variables.

2.1 Univariate Normal

Consider the case where there is only one continuous characteristic X, which follows a normal distribution. The probability distribution function for good and bad payers are $f(x|G)$ and $f(x|B)$ respectively, where $X_G \sim N(\mu_G, \sigma_G)$ and $X_B \sim N(\mu_B, \sigma_B)$. Thus,

$$\frac{f(x|G)}{f(x|B)} = \frac{\frac{1}{\sigma_G}\exp\left\{\frac{-(x-\mu_G)^2}{2\sigma_G^2}\right\}}{\frac{1}{\sigma_B}\exp\left\{\frac{-(x-\mu_B)^2}{2\sigma_B^2}\right\}} = \frac{\sigma_B}{\sigma_G}\exp\left\{\frac{-1}{2}\left[\frac{(x-\mu_G)^2}{\sigma_G^2} - \frac{(x-\mu_B)^2}{\sigma_B^2}\right]\right\} \overset{\text{set}}{\geq} \frac{Dp_B}{Lp_G} \quad (4)$$

From the last inequality we can deduce that

$$\exp\left\{\frac{-1}{2}\left[\frac{(x-\mu_G)^2}{\sigma_G^2} - \frac{(x-\mu_B)^2}{\sigma_B^2}\right]\right\} \geq \frac{Dp_B\sigma_G}{Lp_G\sigma_B}$$

$$\implies \frac{(x-\mu_G)^2}{\sigma_G^2} - \frac{(x-\mu_B)^2}{\sigma_B^2} \leq -2\log\left(\frac{Dp_B\sigma_G}{Lp_G\sigma_B}\right)$$

We can conclude that the set of good payers are

$$A_G = \left\{x \mid \frac{(x-\mu_G)^2}{\sigma_G^2} - \frac{(x-\mu_B)^2}{\sigma_B^2} \leq -2\log\left(\frac{Dp_B\sigma_G}{Lp_G\sigma_B}\right)\right\} \quad (5)$$

2.2 Bivariate Normal

Consider another case where there are two continuous normally distributed characteristics X_1 and X_2, where $X_{1G} \sim N(\mu_{1G}, \sigma_{1G})$, $X_{1B} \sim N(\mu_{1B}, \sigma_{1B})$, $X_{2G} \sim N(\mu_{2G}, \sigma_{2G})$, $X_{2B} \sim N(\mu_{2B}, \sigma_{2B})$ and $corr(X_1, X_2) = 0$. Therefore, the variance covariance matrix of the good and bad payers are

$$\Sigma_G = \begin{pmatrix} \sigma_{1G}^2 & 0 \\ 0 & \sigma_{2G}^2 \end{pmatrix} \quad \text{and} \quad \Sigma_B \begin{pmatrix} \sigma_{1B}^2 & 0 \\ 0 & \sigma_{2B}^2 \end{pmatrix}$$

The corresponding density function for good payers are

$$f(\mathbf{x}|G) = \frac{1}{2\pi|\Sigma_G|^{\frac{1}{2}}}\exp\left\{\frac{-1}{2}(\mathbf{x}-\mu_\mathbf{G})^T\Sigma_G^{-1}(\mathbf{x}-\mu_\mathbf{G})\right\} \quad (6)$$

We can deduce that

$$\frac{f(x|G)}{f(x|B)} = \frac{\frac{1}{2\pi\sigma_{1G}\sigma_{2G}}\exp\left\{\frac{-1}{2}(\mathbf{x}-\mu_\mathbf{G})^T\Sigma_G^{-1}(\mathbf{x}-\mu_\mathbf{G})\right\}}{\frac{1}{2\pi\sigma_{1B}\sigma_{2B}}\exp\left\{\frac{-1}{2}(\mathbf{x}-\mu_\mathbf{B})^T\Sigma_B^{-1}(\mathbf{x}-\mu_\mathbf{B})\right\}} \quad (7)$$

$$= \frac{\sigma_{1B}\sigma_{2B}}{\sigma_{1G}\sigma_{2G}}\exp\left\{\frac{-1}{2}(\mathbf{x}-\mu_\mathbf{G})^T\Sigma_G^{-1}(\mathbf{x}-\mu_\mathbf{G}) - (\mathbf{x}-\mu_\mathbf{B})^T\Sigma_B^{-1}(\mathbf{x}-\mu_\mathbf{B})\right\} \quad (8)$$

$$\overset{\text{set}}{\geq} \frac{Dp_B}{Lp_G} \quad (9)$$

From the above, we conclude that for the bivariate case, the set of good payers are

$$A_G = \left\{\mathbf{x} \mid \mathbf{x}^T(\Sigma_G^{-1} - \Sigma_B^{-1})\mathbf{x} - 2\mathbf{x}(\Sigma_G^{-1}\mu_\mathbf{G} - \Sigma_B^{-1}\mu_\mathbf{B}) \leq \mu_\mathbf{B}^\mathbf{T}\Sigma_B^{-1}\mu_\mathbf{B} - \mu_\mathbf{G}^\mathbf{T}\Sigma_G^{-1}\mu_\mathbf{G} - 2\log\left(\frac{\sigma_{1G}\sigma_{2G}Dp_B}{\sigma_{1B}\sigma_{2B}Lp_G}\right)\right\} \quad (10)$$

3 Simulation of Data

We use simulated data to illustrate our proposed issue, that of prospective borrowers lying about one or more of their attributes. We first simulated a Bernoulli random variable *GB*, to distinguish whether the borrower is a good or bad payer. Good borrowers repay their loans on time, while bad borrowers default. Bernoulli variable *GB* takes value 1 with probability p, otherwise taking the value zero. Hence, p denotes the probability of being a good borrower. If *GB* is one, we say that the borrower is a good payer and we simulate the corresponding characteristics X_{1G} and X_{2G}, where X_{1G} and X_{2G} are normally distributed with mean μ_{1G} and μ_{2G} and standard deviation σ_{1G} and σ_{2G} respectively. On the other hand, if *GB* is zero, we say that the borrower is a bad payer and we simulate the corresponding characteristics X_{1B} and X_{2B}, where X_{1B} and X_{2B} are normally distributed with mean μ_{1B} and μ_{2B}, and standard deviation σ_{1B} and σ_{2B} respectively. Note X_1 and X_2 are independent, i.e. uncorrelated.

$$X_1 = \mu_1 + \sigma_1 Z_1 \text{ , where } Z_1 \sim N(0,1) \tag{11}$$

$$X_2 = \mu_2 + \sigma_2 Z_2 \text{ , where } Z_2 \sim N(0,1) \tag{12}$$

Since we assume that only bad payers will lie about their information and can only alter their characteristics through X_{2B}, in order to make them indistinguishable from the good payers, we added noise to X_{2B} only. We believe that not all bad payers choose to lie about their characteristics; we introduced another Bernoulli random variable *Noi*, which takes the value one when a particular bad payer lies, otherwise taking the value zero. We equate the probability that a bad payer will lie to pN. We further assume that all bad payers who intend to lie do so by adding a fixed constant amount A to their attribute. For example, a liar might say they earned 10K more than they actually did. As a result,

$$X_{2Bnew} = X_{2B} + Noi \times A \tag{13}$$

Table 1: Table of Parameter Values

μ_{1G}	σ_{1G}	μ_{1B}	σ_{1B}	μ_{2G}	σ_{2G}	μ_{2B}	σ_{2B}	D	L	p	pN	A
8	1	4	1	8	1	6	1	5	2	0.5	0.9	$\frac{2}{0.9}$

4 Results

The set condition in Equation 5 and Equation 10 can be applied to do prediction if $f(\mathbf{X}|G)$ and $f(\mathbf{X}|B)$ are normally distributed. However, our data is not in fact normal, rather a mixture of normals. Next we see if our data nonetheless appear normal enough that a busy credit quant might model it as normal. A proportion of lies is added to X_{2B} as described in section 3. To ensure the plausibility of an incorrect normal assumption for X_{2Bnew}, we generated a histogram and a Q-Q plot for X_{2Bnew}, and performed Anderson-Darling (Stephens, 1986; Thode, 2002, Sec. 5.1.4) and Cramer-von Mises (Stephens, 1986; Thode, 2002, Sec. 5.1.3) normality tests to show that the possibility of X_{2Bnew} still follows a normal distribution cannot be rejected. The Q-Q plot in Figure 1 and the two test results in Table 1 ensured that X_{2Bnew} is sufficiently close to a normal distribution to fool many observers.

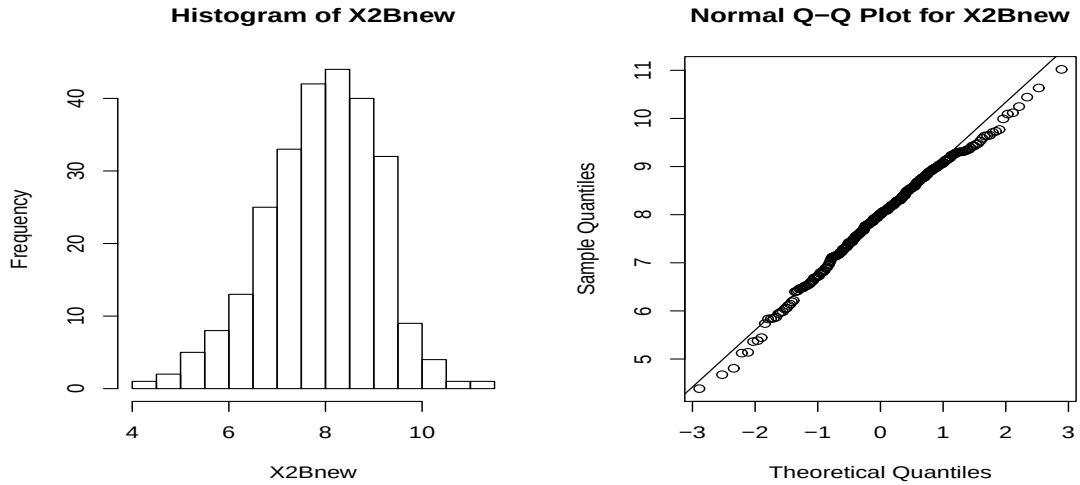

Figure 1: Histogram and Q-Q Plot of X_{2Bnew}

Table 2: Test of Normality

Anderson-Darling Test	Cramer-von Mises Test
p-value = 0.1011	p-value = 0.1338

Using the parameters of Table 1, we simulated 500 datasets each with 500 borrowers and applied the set conditions in Equation 5 and Equation 10 to find the accuracy of the prediction. Note that, in an attempt to keep historical computation representative, classes may have relatively few members. The accuracy of classification was defined as $\frac{(GG+BB)}{\text{Total number of borrowers}}$, where GG is the number of predicted good payers given that they are truly good payers, and BB is the number of predicted bad payers given that they are truly bad payers. We computed the average accuracy of the 500 datasets. From the results shown in Table 3, there is evidence that using only one characteristic to do prediction is on average 1.84% better that using two characteristics.

In fact X_{2Bnew} is not normally distributed; we can find the true distribution of X_{2Bnew} and insert that into discriminant analysis to obtain the correct results. Through the way we generated X_{2Bnew}, we can deduce that

$$f(x_{2Bnew}) = (1-pN)\frac{1}{\sqrt{2\pi\sigma_{2B}^2}}\exp\left\{\frac{-(x-\mu_{2B})^2}{2\sigma_{2B}^2}\right\} + pN\frac{1}{\sqrt{2\pi\sigma_{2B}^2}}\exp\left\{\frac{-(x-\mu_{2Bnew})^2}{2\sigma_{2B}^2}\right\}, \tag{14}$$

where μ_{2Bnew} equals $\mu_{2B}+A$.

Since X_{1B} and X_{2Bnew} are independent and applying inequality 3, we found that the correct set condition is

$$A_G = \left\{ \mathbf{x} \left| \frac{\frac{1}{\sigma_{1G}\sigma_{2G}} \exp\left\{\frac{-(x_1-\mu_{1G})^2}{2\sigma_{1G}^2}\right\} \exp\left\{\frac{-(x_2-\mu_{2G})^2}{2\sigma_{2G}^2}\right\}}{\frac{1}{\sigma_{1B}\sigma_{2B}} \exp\left\{\frac{-(x_1-\mu_{1B})^2}{2\sigma_{1B}^2}\right\} \left[(1-pN)\exp\left\{\frac{-(x-\mu_{2B})^2}{2\sigma_{2B}^2}\right\} + pN \exp\left\{\frac{-(x-\mu_{2Bnew})^2}{2\sigma_{2B}^2}\right\}\right]} \geq \frac{Dp_B}{Lp_G} \right\} \tag{15}$$

Using the above inequality, the correct accuracy of using two characteristics is 97.50%, which is 0.04% better than using only one characteristic. This coincides with our general theory that using more information produce better performance. Note that the incorrect assumption of normality did not cause the 1-attribute classifier any significant degradation in accuracy: Equation 15 gives an accuracy of 97.5%, while the significant simpler Equation 5 gives an accuracy of 97.46%. According to those figures, using two variables only show minor improvements on the accuracy of prediction, compared against using only one variable. Even here, it is likely wise to use only one variable to lower the cost and faster the analyzing process.

Table 3: Table of Results

Attribute	X_1 only	X_1 and X_2	X_1 and X_2 (Corrected Model)
Accuracy	97.46%	95.62%	97.50%

5 Conclusions

Discriminant analysis is a recognized method that has been widely used in credit scoring. In this paper, we showed that using smaller datasets to predict good or bad borrowers are more reasonable and produce more accurate results due to the fact that people misused prediction models in the classification process. More effort should be expended on examining the data to detect fraud and on whether the prediction models are statistically valid. The correct usage of credit scoring models can effectively and efficiently fulfill the needs of loan request and can maximize the profit from the lending of money.

REFERENCES

Anderson, R., (2007) *The Credit Scoring Toolkit*, Oxford University Press, USA.

Stephens, M.A. (1986) “Tests based on EDF statistics”. In: D‘Agostino, R.B. and Stephens, M.A., eds., *Goodness-of-Fit Techniques*, Marcel Dekker, New York

Thode Jr., H.C. (2002) *Testing for Normality*, Marcel Dekker, New York.

Thomas, L.C., Edelman, D.B., Crook, J.N., (2002) *Credit Scoring and its Application*, SIAM, Philadelphia, USA.

Thomas, L.C., (2000) “A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers,” *International Journal of Forecasting*, 16, 149–172.