Probit Regression over Company Attributes and Bankruptcy Data

Jianyu Lin, Honglin Fu
Department of Applied Mathematics and Statistics
Johns Hopkins University
Baltimore, MD 21218
jlin153@jh.edu
hfu10@jh.edu

Abstract

With the advancement of forecasting techniques and accumulation of big data, bankruptcy forecasting has become a hot topic and attracts a lot of researchers. There have been many machine learning methods studies towards this topic, such as Neural Networks and Support Vector Machines. Although these machine learning methods produce good forecast results, they are not intuitive in financial and economic sense. In this paper, we introduce a Bayesian statistic method. This method helps us to select the most influential factors so that we can have a clear understanding on which financial attributes are affecting the bankruptcy. Moreover, by refining our model, we remove some attributes and confirm the significant ones. Finally, we offered an interpretation of the model and an explanation on how the attributes contribute to the bankruptcy.

1 Introduction

The COVID-19 pandemic has greatly interrupted the economic stability worldwide. The unemployment rate has surged, and retail sale has plunged. Massive manufacture shutdowns are accompanied by large decrease in consumer spending have worsen the situation for business. Most major economies in the world experience negative GDP growth during the pandemic. Although, the economy has bounded back since the introductions of vaccine and government stimulation policy, the world economy clearly enters a stage of slow growth and uncertainty. According to study of Harvard Business School (Wang, Yang, Iverson, Kluender, 2020). There is an increase in the cases of large corporation filling bankruptcies but a decrease in the cases of small and medium size companies. The difference can be explained by the fact that large corporation uses bankruptcy as a protection, but small business perceives it as a last measure and avoid filling. This unique dynamic makes the study of uncertainty of bankruptcy and factors driving the event more crucial in the contemporary setting.

The assets of bankrupt companies are evaluated and used to cover the debt as much as possible. Thus, it is in the interest of its shareholders and creditors to predict the bankruptcy. For the company, to anticipate a bankruptcy before such an event helps the management team steer the company back to the right direction or fill a bankruptcy case to secure a second chance to the credit. For the creditor, after the financial crisis, financial institution requires information of risks of bankruptcy to price the firm assets and credit derivatives. Traditional methods of ratio analysis are proved to be useful, but the technique is outdated. The limitation is that it fails to incorporate a large number of factors, a problem in the era of big data.

In the recent years, the data existing in the world has increased exponentially. With the adoption of online business service, more and more factors are used to evaluate the financial status of companies. This creates an ideal application environment for machine learning method. According to the article,

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

Review of Bankruptcy Prediction Using Machine Learning and Deep Learning Techniques (Qu, Quan, Lei, Shi, 2019), common machine learning methods used to forecast bankruptcy are Neural Networks, Multivariant Discriminant Analysis, and Support Vector Machine. These methods achieve good prediction accuracy but lacks the interpretability of the model. To have a better understanding of how the factors are linked to the prediction, researchers have turned their focus to the Bayesian statistics method.

2 Data

We use the same dataset as the research paper "Financial Data Analysis Using Expert Bayesian Framework For Bankruptcy Prediction". The dataset contains bankruptcy information of Polish companies, collected from year 2000 to year 2012.

In the dataset, there are 5600 data entries and 64 attributes. For the attributes, there are accounting ratio such as net profit to total assets and total liabilities to total assets. A complete list of variables will be provided in supplementary material. We randomly divide the 5600 samples into training set (3000 samples) and test set (2600 samples). We will train our model on the test set and calculate the prediction error over the test set. A brief data summary of our training and test sets are as below:

Set	Bankrupted Companies	Non-bankrupted Companies	Total
Training set	62	2938	3000
Test set	51	2549	2600

3 Model

Let X be the attributes, and Y be the class whether the company is finally bankrupted. $Y_i = 1$ indicates the i-th company finally bankrupted, and $Y_i = 0$ means the i-th company did not bankrupt. Since we only have two discrete values of Y ($Y = 0$ and $Y = 1$), we cannot directly linearly regress Y against X. To solve this problem, we introduce a latent variable Z, and use the binary probit model on our dataset.

$$Z_i = \beta^\top X_i + \epsilon_i$$
$$Y_i = \delta_{(g,\infty)}(Z_i) = \begin{cases} 1 \text{ if } Z_i > g \\ 0 \text{ otherwise} \end{cases}$$

We assume linear model Z over X. Since we can use g to adjust the scale, we directly assume the error $\epsilon_i \sim N(0, 1)$. Our goal is to solve the parameter β and g on the training set.

4 Method

The latent variable Z is unknown. To solve the parameter β and g in our probit model, we need to use Gibbs sampler. We first calculate the full conditional of β, Z, and g.

4.1 Full conditional distribution of β

Let the prior distribution of β be

$$\beta \sim \text{multivariate normal}(0, n(X^\top X)^{-1})$$

According to our model, β does not depend on X, Y, g. Then the posterior distribution of β given Z is a multivariate normal distribution with

$$\mathbb{E}(\beta|Z) = \frac{n}{n+1}(X^\top X)^{-1}X^\top Z$$
$$\text{Var}(\beta|Z) = \frac{n}{n+1}(X^\top X)^{-1}$$

That is,

$$\beta|Z \sim \text{multivariate normal}(\frac{n}{n+1}(X^\top X)^{-1}X^\top Z, \frac{n}{n+1}(X^\top X)^{-1})$$

4.2 Full conditional distribution of Z

If only condition Z over X and β, then according to the model, we have

$$Z_i|X \sim \text{multivariate normal}(\beta^\top X_i, 1)$$

However, as for the full conditional, we have further limitation over Z_i once Y_i and g is given. If $Y_i = 1$, then according to the model, we must have $Z_i > g$. If $Y_i = 0$, then according to the model, we must have $Z_i \leq g$. So $Z_i|\beta, X, Y, g$ has a truncated normal distribution

So the full conditional distribution of Z is given by

$$Z_i|\beta, g, X_i, Y_i = 1 \sim \text{truncated normal}(\beta^\top X_i, 1, a = g, b = \infty)$$
$$Z_i|\beta, g, X_i, Y_i = 0 \sim \text{truncated normal}(\beta^\top X_i, 1, a = -\infty, b = g)$$

Where a and b is the minimum value and maximum value of the truncated normal distribution, respectively.

4.3 Full conditional distribution of g

Given Z and Y, we know that g must be greater than all Z_i such that $Y_i = 0$. At the same time, g must be smaller than all Z_i such that $Y_i = 1$.

Denote

$$s = \max\{Z_i|Y_i = 0\}, t = \min\{Z_i|Y_i = 1\}$$

So we have $s < g < t$. In fact, g can be any number in the interval (s, t), because any $g \in (s, t)$ provides the same separation for $Y_i = 0$ and $Y_i = 1$.

Hence the full conditional distribution of g is given by

$$\begin{aligned} g|Z, Y &= \text{Uniform}(s, t) \\ &= \text{Uniform}(\max\{Z_i|Y_i = 0\}, \min\{Z_i|Y_i = 1\}) \end{aligned}$$

5 Implementation

Observe that the we have 64 attributes for each companies in our dataset. Too much attributes might lead to correlation. Also, our goal is to predict whether a company would bankrupt using the probit model. In real life, it is inconvenient, or even impossible, to collect all the 64 attributes of certain company. So to improve the accuracy as well as the convenience, we want to reduce the size of attributes employed in the model. We will first run the original model (all 64 attributes included), and see how we can make improvement.

5.1 Original model

Given the full conditional distribution of β, Z, g, we can now run the Gibbs Sampler over the training set.

In our implementation, we first set initial values

$$Z = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^N$$
$$g = 1$$

N is the size of training set, N = 3000. We run the Gibbs Sampler algorithm for 5000 iterations, save

the value of β and g in each iteration. We can get the posterior distribution of g through the below histogram of g. The posterior mean of g is $g = 1.241$

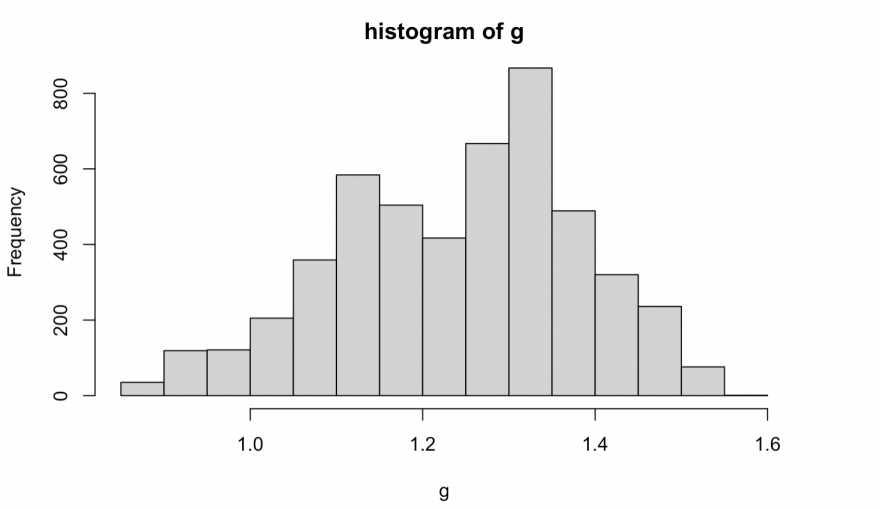

Figure 1: Histogram for parameter g

We also find the 95% posterior confidence interval for each entry of β

95 Posterior Interval			95 Posterior Interval		
Index of Attribute	Lower bound	Upper bound	Index of Attribute	Lower bound	Upper bound
1	-28.01	10.02	33	5.00	13.82
2	-7.32	67.60	34	0.57	2.63
3	-0.30	0.58	35	-0.38	0.27
4	-2.13	1.48	36	-0.34	1.23
5	-0.01	0.72	37	-1.74	0.41
6	-0.56	0.34	38	-9.67	12.25
7	-480326.28	593920.62	39	-11.08	-3.62
8	-5.36	11.23	40	0.22	4.17
9	-0.99	0.49	41	-0.24	0.06
10	-10.19	50.13	42	-73.55	32.58
11	-0.86	1.24	43	-64499.36	18594.60
12	-1.26	2.14	44	-8926.13	30954.42
13	-403.24	1076.65	45	-16.66	1.83
14	-598425.50	479199.47	46	-5.84	-0.30
15	-0.34	0.51	47	-4.32	12.74
16	-6.26	7.04	48	1.56	5.12
17	-11.65	5.52	49	-32.70	71.13
18	-6598.90	9293.94	50	-2.57	-0.57
19	-92.60	28.21	51	-0.99	0.45
20	-11180.01	38778.30	52	-60.39	-1.21
21	-75.95	2.74	53	0.18	4.18
22	-3.58	-0.37	54	-67.66	-18.82
23	-15.64	32.92	55	-0.29	0.14
24	-0.02	0.39	56	2.86	9.81
25	-0.68	0.57	57	-0.27	0.39
26	-5.57	6.40	58	-5.77	3.19
27	-43.39	4.96	59	-1.03	0.29
28	17.49	66.48	60	-2.71	0.48
29	-0.02	0.36	61	-0.95	-0.02
30	-12.29	-2.08	62	-5.36	14.81
31	1.45	20.08	63	-18.60	-6.60
32	-4.38	54.85	64	-6.63	-0.58

A parameter would be considered not significant if 0 belongs to its 95% posterior confidence interval. Therefore, we remove all the attributes with parameters not significant. Then we obtain 18 attributes below whose parameters are significant. So these attributes would have a considerable influence on whether the company would bankrupt. The below table shows the posterior mean of the β parameters for the selected 18 attributes.

Index of Attribute	posterior mean	Index of Attribute	posterior mean
22	-2.01	48	3.66
28	40.56	50	-1.58
30	-7.47	52	-29.61
31	10.85	53	1.87
33	10.08	54	-42.01
34	1.61	56	6.46
39	-7.3	61	-0.51
40	2.49	63	-13.84
46	-2.81	64	-3.14

Now we can use the posterior mean of β and posterior mean of $g = 1.241$ to make predication over the test set. We get 67 incorrect prediction over the test set. The error rate is 0.0258.

Even though the error rate is very low, our model performs well in correctly predicted a company that would not bankrupt, but our model performs poor in make correct prediction on the company that bankrupted. That is, even though the overall performance of our model is great, our model has high false negative error – among the 51 bankrupted companies in the test set, only 2 of them are correctly detected through the prediction. Therefore, we would like to further improve our model in the following subsection.

5.2 Improvement

By observing the variables, we notice that there exists great correlation between different attributes. For example, Attribute2 (total liabilities/total assets) and Attribute17 (total assets/total liabilities) are inversely proportional to each other. So even though in previous implementation, we already remove the attributes variables, the existence of too many attributes while implementing the Gibbs Sampler would cause inaccuracy of the significant attributes. Therefore, we run the algorithm again over the training set, only using the 18 selected attributes above.

In this time, we run the Gibbs Sampler for 10000 iterations, save the value of β and g in each iteration. We can get the posterior distribution of g through the below histogram of g. The posterior mean of g is $g = 3.506$

The below table presents the posterior mean, lower and upper bounds of 95% posterior confidence interval of the parameter β for the 18 selected attributes. We can notice the 95% posterior confidence interval of Attribute 30, Attribute 31, Attribute 53, Attribute 61 contains 0. So the parameters of these four attributes are not significant. Hence, we would again remove the 4 attributes from our model.

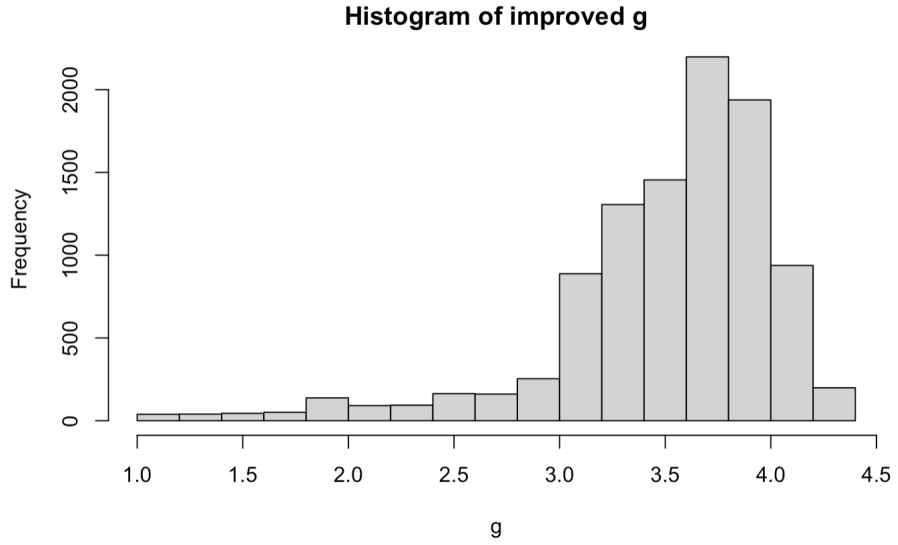

Figure 2: Histogram for improved parameter g

Index of Attribute	Posterior lower bound	Posterior upper bound	Posterior mean
22	-3.52	-0.55	-2.25
28	12.18	48.61	30.74
30	-3.38	0.05	-1.43
31	-0.42	4.58	2.06
33	2.25	14.13	9.44
34	0.22	1.55	0.98
39	-8.38	-2.26	-5.64
40	0.25	3.03	1.98
46	-4.54	-0.38	-2.84
48	0.80	4.53	3.01
50	-1.24	-0.09	-0.73
52	-6.00	-0.07	-2.27
53	-0.42	1.94	0.55
54	-46.10	-10.15	-28.68
56	2.35	8.40	5.85
61	-0.65	0.01	-0.28
63	-16.67	-2.64	-11.54
64	-4.61	-0.10	-2.22

Then we can finally obtain our probit regression model

$$\begin{aligned} Z_i = &- 2.25\text{Attr22} + 30.74\text{Attr28} + 9.44\text{Attr33} + 0.98\text{Attr34} \\ &- 5.64\text{Attr39} + 1.98\text{Attr40} - 2.84\text{Attr46} + 3.01\text{Attr48} \\ &- 0.73\text{Attr50} - 2.27\text{Attr52} - 28.68\text{Attr54} + 5.85\text{Attr56} \\ &- 11.54\text{Attr63} - 2.22\text{Attr64} + \epsilon_i \\ Y_i = &\delta_{(3.506,\infty)}(Z_i) \end{aligned}$$

Apply this probit regression model over the test set, we get 61 incorrect prediction over the test set. The error rate is 0.0235. The overall error rate is improved. Also, we have big improvement in detecting bankrupted company through prediction. Using the new probit model, 10 out of the 51 bankrupted companies in the test set have been correctly detected through prediction. So the success rate of correctly predict the company would finally bankrupt among the 51 bankrupted companies increases by 5 times.

6 Analysis

In this section we will discuss the attributes that have great influence over the bankruptcy of the company. That is, we would talk about the attributes in the model that has big absolute value. Attribute 28, Attribute 33, Attribute 54, and Attribute 63 have absolute value near 10 or greater than 10. We want to analysis the relationship between these four attributes and company bankruptcy.

6.1 Attribute 28: working capital / fixed assets

The coefficient of **working capital / fixed assets** is 30.74, so our model indicates the greater the ratio **working capital / fixed assets** is, the company is more tended to bankrupt. Company sometimes need additional working capital due to external factors (e.g. market fluctuation). Under such situation, sometimes the company need to turn fixed assets to working capital to solve the needs. The additional working capital needed is always in proportion to the current working capital. If the ratio is large, then relative to the working capital, less fixed assets can be transformed. So the company is less stable towards external factors. Hence the company has higher possibility to bankrupt.

6.2 Attribute 33: operating expenses / short-term liabilities

The coefficient of **operating expenses / short-term liabilities** is 9.44. So our model suggests that the greater the ratio **operating expenses / short-term liabilities** is, the more likely the company would bankrupt. If the ratio is large, the money invested in operating exceeds the debts the company needs to pay within certain time period, which would lead to bankruptcy.

6.3 Attribute 54: constant capital / fixed assets

The coefficient of **constant capital / fixed assets** is -28.68. So our model suggests the greater the ratio **constant capital / fixed assets** is, the less possible the company would bankrupt. Constant capital is the value of goods and materials required to produce a commodity. The greater the ratio **constant capital / fixed assets** is, the more resources of the company is utilized in production. Thus, a company with this ratio high has great production utilization. Hence the company is less likely to bankrupt.

6.4 Attribute 63: sales / short-term liabilities

The coefficient of **sales / short-term liabilities** is -11.54. So our model indicates that the greater the ratio **sales / short-term liabilities** is, the company is less likely to bankrupt. The ratio sales / short-term liabilities can be understood as the ratio between the money earned in one year (or some short-term time) divided by the debt the company needed to pay in that short-term time. So if the ratio is large, it means that within that time period, the company is able to earn enough to cover the debt. So the larger the ratio, the more likely that company can decrease or even erase its liabilities – therefore the company is less likely to bankrupt due to huge liability.

7 Summary

To summarize, our model points out that large value of Attributes 28, 33, 34, 40, 48, 56, would make the company easier to bankrupt; large value of Attributes 22, 39, 46, 50, 52, 54, 63, 64 would make the company difficult to bankrupt. Company can use these variable to measure whether itself is dangerous for bankruptcy. On the other hand, we need to admit that the probit model we finally obtained is very theoretical in the sense of Bayesian. So some variable in the model may lack of meaning in financial theory. Still, our model can be a helpful reference for company to measure itself, seek out the problems, and make plan for improvement.

References

[1] Mukeri, A., Shaikh, H., amp; Gaikwad, D. D. P. (2020, October 30). Financial Data Analysis using expert bayesian framework for bankruptcy prediction. arXiv.org. Retrieved December 19, 2021, from https://arxiv.org/abs/2010.13892.

[2] Bankruptcy and the COVID-19 crisis - harvard business school. (n.d.). Retrieved December 19, 2021, from https://www.hbs.edu/ris/Publication.

[3] Qu, Yi, et al. "Review of Bankruptcy Prediction Using Machine Learning and Deep Learning Techniques." Procedia Computer Science, Elsevier, 31 Dec. 2019, https://www.sciencedirect.com/science/article/pii/S1877050919320769.