Credit Risk Model Based on Logistic Regression and Weight of Evidence

Xiang YANG\textsuperscript{1,a,*}, Yongbin ZHU\textsuperscript{1,b}, Li YAN\textsuperscript{1,c}, Xin WANG\textsuperscript{1,d}

\textsuperscript{1}School of Engineering Technology, Honghe University, Mengzi, 661100, China

\textsuperscript{a}email: 806464637@qq.com, \textsuperscript{b}email: 410239127@qq.com, \textsuperscript{c}email: 904382815@qq.com, \textsuperscript{d}email: 2535563360@qq.com, \textsuperscript{*}corresponding author

\textbf{Keywords:} Credit Risk, Logistic Regression, Weight of Evidence, Scorecard

\textbf{Abstract.} Many techniques have been used to build credit risk model. Among them, logistic regression is a more appropriate technique due to its desirable features (e.g., interpretability and prediction accuracy). In this paper, to implement credit risk assessment quickly, a method for constructing credit risk model (in the form of a scorecard) based on logistic and weight of evidence is proposed.

\textbf{Introduction}

Effective management of credit risk is important for a loans institution. A loans institution always willing to lend to creditworthy customers (called good customers), and reject the loan application of customers with poor credit (called bad customers). To distinguish good customers and bad customers using application information, many techniques have been used to build credit risk model: linear discriminant analysis \cite{1}, artificial neural network \cite{1} \cite{3}, \textit{k}-nearest neighbor \cite{2}, classification tree \cite{3}, logistic regression \cite{3} \cite{4} and so on.

Although these techniques can provide good discrimination, logistic regression is a more appropriate technique to build credit risk model due to its distinctive features (interpretability, prediction accuracy and so on \cite{5}). Especially, regulators require that loans institutions give reasons for rejecting a loan application; in this regard, logistic regression is satisfactory because of its interpretability.

Simple logistic regression model is not convenient in practical applications because of a lot of computing. Therefore, the credit risk model (in the form of a scorecard) based on logistic regression and weight of evidence is proposed to distinguish good customers and bad customers quickly.

\textbf{Logistic regression on credit data}

Logistic regression used to divide multidimensional vectors into two categories. On credit data, logistic regression form is given by

\begin{equation} p_n = \frac{\exp(\beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,2} + \ldots + \beta_k x_{n,k})}{1 + \exp(\beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,2} + \ldots + \beta_k x_{n,k})} \end{equation}

where

- $p_n$ = the probability of costumer n’s class label is bad (true or 1)
- $x_{n,i}$ = the value of costumer n on characteristic $i$
- $\beta_i$ = the coefficient of the model

In logistic regression, the value of class label for a costumer is discrete; the values of $\beta$ can be estimated by the method of maximum likelihood estimation. The likelihood function form is given by

\begin{equation} L = \prod_{j=1}^{N} \frac{\exp(\beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,2} + \ldots + \beta_k x_{n,k})}{1 + \exp(\beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,2} + \ldots + \beta_k x_{n,k})} \end{equation}

where $N$ is the number of costumers. Then, the values of $\beta$ can be obtained by solving the following equations.
A good logistic regression model should have significant linear relationship between the tuple of \((x_1, x_2, \ldots, x_k)\) and \(\text{logit}(p)\) (i.e., \(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k\)). Therefore, set the hypotheses as

\[
\begin{align*}
H_0: & \quad \beta_1 = \beta_2 = \ldots = \beta_k = 0 \\
H_1: & \quad \beta_1 = \beta_2 = \ldots = \beta_k = 0 \text{ is false}
\end{align*}
\]

The test statistic adopted is the log likelihood ratio, and the form is given by

\[
G^2 = 2 \ln \frac{L}{L_0}
\]

where \(L_0\) is the value of \(L\) while \(H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0\). The \(G^2\) obeys chi-square distribution with \(df = k\). After calculation, null hypothesis will be rejected if the p-value is less than significance level, and the linear relationship between the tuple of \((x_1, x_2, \ldots, x_k)\) and \(\text{logit}(p)\) is considered to be significant.

Analogously, a good logistic regression model should have significant linear relationship between \(x_i\) and \(\text{logit}(p)\). Therefore, set the hypotheses: \(H_0: \beta_i = 0\), and \(H_1: \beta_i \neq 0\). The test statistic is the Wald that obeys chi-square distribution with \(df = 1\) and the form is given by

\[
\text{Wald} = \left(\frac{\beta_i}{S_{\beta_i}}\right)^2
\]

where \(S_{\beta_i}\) is the standard error of \(\beta_i\). The linear relationship between \(x_i\) and \(\text{logit}(p)\) is considered to be significant if the p-value is less than significance level.

**Weight of evidence on credit data**

The value of weight of evidence (woe) can maximize the predictive ability of variable [6]. For example, the values of characteristic \(x\) are divided into three bins: bin 1, bin 2 and bin 3; then, bin 1’s woe is given by

\[
\ln \left(\frac{gr}{br}\right) \times 100
\]

where \(gr\) is the number of good costumers (while the value on \(x\) belongs to bin 1) is divided by the number of all good costumers; \(br\) is the number of bad costumers (while the value on \(x\) belongs to bin 1) is divided by the number of all bad costumers. Need to be explained, the reason of binning is to meet scorecard’s needs; a costumer’s value on a characteristic must belong to a bin, and the credit is measured by the bin’s score.

**Empirical experiment**

For modeling, a credit dataset (to protect privacy, the dataset was processed specially) from University of California at Irvine Machine Learning Repository is applied. The dataset include 690 costumers’ information; 15 characteristics (A1−A15; A1, A4, A5, A6, A7, A9, A10, A12 and A13 are discrete; A2, A3, A8, A11, A14 and A15 are continuous) and a class label (A16, “+”: good costumers or “−”: bad costumers). In this section, the software of Clementine was used for modeling process and analysis.

In 690 samples, there are 37 samples (5%) with incomplete information; therefore, the 37 samples were discarded; in the remaining 653 samples, 296 are good, and 357 are bad.

The characteristics that have little effect on the prediction for class label must be discarded. Therefore, the adopted strategies is: (1) a discrete characteristic will be discarded if the number of samples with a special value on this characteristic divided by the number of all samples is greater than 90%; (2) a discrete characteristic will be discarded if the number of arbitrary two sample values on this characteristic are not the same divided by the number of all samples is greater than 95%; (3) a continuous characteristic will be discarded if its variance coefficient is less than 0.1; (4) a characteristic will be discarded if the p-value (based on the hypothesis testing: \(H_0: \beta_i = 0\))
characteristic is not related to the class label, \( H_1 \): the characteristic is related to the class label) is greater than 0.05. So, A13 is discarded according to strategy no. 1; A1 and A12 were discarded according to strategy no. 4.

To reduce the computational intensity, the characteristics that are redundant for predicting class label must be discarded. Therefore, the adopted strategy is to retain one of the two if the two continuous characteristics are related (the p-value is less than 0.05). So, A2, A3 and A11 are discarded. Note: to measure the correlation between continuous characteristic 1 and 2, the hypothesis is:

\[
H_0: \text{characteristic 1 and 2 are not related, } H_1: \text{characteristic 1 and 2 are related.}
\]

According to the features of a scorecard, a costumer’s value on a characteristic must belong to a bin. Therefore, the adopted strategy is: (1) the number of bins should be appropriate; (2) each bin of a characteristic has special effect on class label distribution. The weight of each bin is represented by woe that can be calculated by Equation 6. After processing, the rest of characteristics, their bins and woes of bins are shown in Tab. 1.

<table>
<thead>
<tr>
<th>Bins</th>
<th>Woes</th>
<th>Bins</th>
<th>Woes</th>
<th>Bins</th>
<th>Woes</th>
</tr>
</thead>
<tbody>
<tr>
<td>A6 ff</td>
<td>-162.791</td>
<td>A8 &lt;1.21</td>
<td>-79.608</td>
<td>A4 l</td>
<td>148.462</td>
</tr>
<tr>
<td>x</td>
<td>179.681</td>
<td>&gt;=1.21</td>
<td>85.595</td>
<td>u</td>
<td>18.337</td>
</tr>
<tr>
<td>aa, m, c, w</td>
<td>-2.319</td>
<td>A9 f</td>
<td>-257.824</td>
<td>y</td>
<td>-67.879</td>
</tr>
<tr>
<td>e, q, r, cc</td>
<td>85.947</td>
<td>t</td>
<td>155.232</td>
<td>A5 g</td>
<td>18.337</td>
</tr>
<tr>
<td>i, d, k, j</td>
<td>-82.668</td>
<td>A10 f</td>
<td>-88.950</td>
<td>gg</td>
<td>148.462</td>
</tr>
<tr>
<td>A7 ff</td>
<td>-156.182</td>
<td>t</td>
<td>106.977</td>
<td>p</td>
<td>-67.879</td>
</tr>
<tr>
<td>z</td>
<td>128.599</td>
<td>A14 &lt;100</td>
<td>68.270</td>
<td></td>
<td></td>
</tr>
<tr>
<td>v, bb, o, n</td>
<td>-8.706</td>
<td>&gt;=100</td>
<td>-31.987</td>
<td></td>
<td></td>
</tr>
<tr>
<td>h</td>
<td>74.126</td>
<td>A15 &lt;500</td>
<td>-45.203</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dd, j</td>
<td>-40.041</td>
<td>&gt;=500</td>
<td>157.367</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

It should be noted: (1) for discrete characteristics with small number of possible values, each possible value regards as a bin; (2) for discrete characteristics with big number of possible values (e.g., A6 and A7), first, each possible value regards as a bin, and calculate the woe of each bin; second, merge bins if the bins' woe are similar (by the method of \(K\) mean clustering); third, recalculate the woe of each bin.

Now, a costumer’s information on a characteristic is replaced by woe. Credit dataset is divided into a training dataset and test dataset randomly; each one includes 50% of samples. The logistic regression after training is shown in Tab. 2.

<table>
<thead>
<tr>
<th>( \beta_0 )</th>
<th>( \beta_4 )</th>
<th>( \beta_6 )</th>
<th>( \beta_8 )</th>
<th>( \beta_9 )</th>
<th>( \beta_{10} )</th>
<th>( \beta_{15} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1475</td>
<td>-0.01673</td>
<td>-0.007306</td>
<td>-0.005736</td>
<td>-0.009532</td>
<td>-0.006775</td>
<td>-0.008818</td>
</tr>
</tbody>
</table>

From Tab. 2, obviously, 3 characteristics (A5, A7, and A14) are discarded by the model.

By the analysis, the statistics of \(G^2\) is 282.903, and the p-value is very close to 0; the \(Wald\) of A4, A6, A8, A9, A10 and A15 are 11.661, 7.655, 5.308, 64.049, 10.487 and 10.816 respectively, and the p-values are all very close to 0.

Percentage Correctly Classified (PCC) represents the percentage of observations that are correctly classified. PCC is also proposed to measure the logistic regression model goodness of fit. PCC based on the training dataset is shown in Tab.3, and Tab.4 for PCC based on the test dataset.
Tab. 3. PCC based on the training dataset

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Observed</th>
<th>PCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>140</td>
<td>17</td>
</tr>
<tr>
<td>1</td>
<td>15</td>
<td>158</td>
</tr>
<tr>
<td>Overall Percentage</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Tab. 4. PCC based on the test dataset

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Observed</th>
<th>PCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>120</td>
<td>29</td>
</tr>
<tr>
<td>1</td>
<td>21</td>
<td>153</td>
</tr>
<tr>
<td>Overall Percentage</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Based on the above analysis, it can be seen that the logistic regression is good.

To implement scorecard, the score of each bin must be determined. In banking industry, Equation 7 is often used to calculate a customer’s credit score.

\[
\text{Score} = \text{Offset} + \text{Factor} \times \ln(\text{odds})
\]  

(7)

Where \(\text{odds} = \) the number of good customers is divided to the number of bad customers on \(\text{Score}\). Usually, the \(\text{odds}\) should have doubled if the \(\text{Score}\) adds some. Let \(\text{odds} = 30/1\) while \(\text{Score} = 500\), and \(\text{odds}\) will have doubled while \(\text{Score}\) for each additional 50; then, a customer’s \(\text{Score}\) is shown by Equation 8.

\[
\text{Score} = 254.6553 + 72.1348 \times \ln(\text{odds})
\]  

(8)

Because \(\logit(p) = -\ln(\text{odds})\), Equation 9 is obtained by Equation 8 and the logistic regression.

\[
\text{Score} = \sum \left[ \frac{254.6553}{6} - 72.1384 \times \left( \frac{0.1475}{6} + \beta_i \times \text{woe} \right) \right]
\]  

(9)

According to Equation 9, the score of a bin on \(A_i\) is

\[
\frac{254.6553}{6} - 72.1384 \times \left( \frac{0.1475}{6} + \beta_i \times \text{woe} \right)
\]

(10)

where, the “\(\text{woe}\)” is the bin’s woe. According to Equation 10, each bin’s score is shown in Tab. 5.

Tab. 5. Bin’s scores

<table>
<thead>
<tr>
<th>Bins</th>
<th>Scores</th>
<th>Bins</th>
<th>Scores</th>
<th>Bins</th>
<th>Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>A6ff</td>
<td>-45</td>
<td>A8</td>
<td>&lt;1.21</td>
<td>A4</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>76</td>
<td></td>
<td>220</td>
</tr>
<tr>
<td>x</td>
<td>135</td>
<td></td>
<td>&gt;=1.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>u</td>
<td></td>
<td>63</td>
</tr>
<tr>
<td>aa, m, c, w</td>
<td>39</td>
<td>A9</td>
<td>-137</td>
<td>y</td>
<td>-41</td>
</tr>
<tr>
<td>e, q, r, cc</td>
<td>86</td>
<td>t</td>
<td>147</td>
<td>A15&lt;500</td>
<td>12</td>
</tr>
<tr>
<td>i, d, j</td>
<td>-3</td>
<td>A10</td>
<td>-3</td>
<td>&gt;=50</td>
<td>141</td>
</tr>
<tr>
<td></td>
<td></td>
<td>t</td>
<td>93</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The scorecard was analyzed using K-S index method on credit dataset, and the analysis result on credit dataset is given by Fig. 1.
In Fig. 1, the horizontal axis represents Score, and the vertical axis represents cumulative ratio. The under curve is to good customer, and the up curve is to bad customer; the two curves are far apart, and it indicates that the scorecard model is good. After calculating, the distance between the two curves is furthest on the Score of 170; on Score of 170, the cumulative ratio of bad customers is 82.633%, and 6.757% for good customers; it means that, if accept a customer with over 170 and reject a customer with less 179, will reject 82.633% of bad customers and 6.757% of good customers. The result shows that the scorecard model (Tab. 3) is good, and the critical value of Score is 170 (note: the Score range is [-206, 812]).

**Conclusion**

In this paper, a model construction method for evaluating credit risk is proposed.

For credit risk evaluation, logistic regression is a more appropriate technique due to its desirable features. A scorecard model can evaluate a costumer’s credit risk quickly at front-end of business. Because of the features of a scorecard (i.e., a costumer’s value on a characteristic must belong to a bin), characteristics in credit dataset was binned; replace the value with bin’s woe. The logistic regression model was proposed based on the training dataset; the analysis result shows that the model is good. Based on the logistic regression model, the scores of bins were determined; the score range and the critical value of score were determined too; the analysis result shows that the scorecard model is good.

To protect privacy, the credit dataset was processed specially; in practical applications, a data analyst can do more works using semantic analysis; for example, obtaining more interesting conclusion, understanding dataset better, guiding business, improving model.

**Acknowledgement**

This work was financially supported by Scientific Research Foundation Project of Honghe University of China (Project No. XJ15Y21).

**References**


