Quantifying credit-scoring performance

G. Verstraeten & D. Van den Poel

Department of Marketing, Ghent University, Belgium

Abstract

Recently, the consumer credit industry has experienced a sizeable growth, where scoring techniques have grown to outperform the traditional, judgmental manner of assessing credit risk. Using the data of a Belgian direct-mail company offering consumer credit, the authors have shown a clear improvement in comparison with its current credit evaluation system, constructed by an international company specialized in consumer credit scoring, and identify the size of this performance increase due to population drift versus model improvement. Considering the crucial impact of the accuracy of the score on the cost side (lowered credit risk) as well as on the revenue side (increased accepted applications), in this study, we highlight the importance of introducing different performance measures to quantify credit risk performance. Hence, instead of reporting predictive performance on the total sample, we have quantified the predictive performance in more detail into a graphical overview that can be used for effective decision making. More specifically, we have augmented the traditional performance measurement by evaluating credit risk as well as credit-scoring profitability in an entangled manner because of the intertwined nature of the outcomes of an adaptation of the current credit-scoring algorithm.

1 Introduction

In this paper, the term ‘credit scoring’ is used as a common denominator for the statistical methods used for classifying applicants for credit into ‘good’ and ‘bad’ risk classes. Using various predictive variables from application forms, external data suppliers and own company records, statistical models, in the industry often termed scorecards, are used to yield estimates of the probability of defaulting. Typically, an accept or reject decision is then taken by comparing the estimated probability of defaulting with a suitable threshold (see e.g. [1]). Considering the widespread use of the statistical scoring techniques in the consumer credit
industry, and considering the longevity of consumer credit-scoring research (see, e.g. [2] for an early application), the literature surrounding customer credit scoring has been growing steadily (for an overview, see, e.g. [3]). In this study, we focus on application scoring (see, e.g. [4] for an overview), i.e., we consider the decision whether or not to grant credit to potential lenders upon application, in contrast to the more recently introduced behavioral scoring (see, e.g. [5]) where the performance of the customer is assessed for decision-making purposes during the lifetime of the relevant credit (e.g. whether the credit limit of a current borrower should be increased). Hence, we focus on the core application within the domain, resulting in a binary classification problem with zero-heavy data.

In the recent credit-scoring literature, the focus has been on benchmarking different credit-scoring algorithms, including statistical and machine learning techniques. The conclusions of this research, however, do not consistently favor one technique over another. In these studies, the problem of credit-score improvement is very often reduced to the mere problem of improving classification accuracy given a set of dependent and independent variables. This leaves improvement in the credit-scoring domain open in two directions where human interactions can matter, namely (i) in terms of the creation of meaningful variables and (ii) in terms of the interpretation and the choice of the optimal credit-score model. Both issues will be addressed in this study.

In this paper, we attempt to improve the current credit-scoring system of a Belgian direct-mail company, and we argue that a sound understanding of the business question and the company data can lead to improvements that are in size at least comparable to the improvements that have been suggested when considering different classification techniques. Hence, as several authors in the domain have stated before, successfully applying data-mining techniques for business purposes cannot simply be reduced to finding the optimal way of relating a dependent variable to a list of independent variables.

The remainder of this paper is structured as follows. In the following section, we will try to lead the reader through the existing credit-score literature, which will hence provide support for the methodology used in this paper for the construction of a new credit score. In Section 3, we will describe briefly the data that was used for this study, and the variables that were created from the data available. Section 4 continues to explain and quantify the improvements that were made to the current credit score, where the improvements will be tested on different possible business objectives. In Section 5, we offer conclusions for our results, and Section 6 completes this paper with some issues for further research.

2 Methodology

2.1 Credit-scoring technique

Recent research in credit scoring has been focused on comparing the performance of different credit-scoring techniques, such as neural networks (ANN), recursive partitioning algorithms (RPA), linear programming (LP), k-nearest neighbour, support vector machines, discriminant analysis, survival
analysis and linear and logistic regression (see, e.g. [6], [7], [8] and [9]). The main conclusions from these efforts are that the different techniques often reach comparable performance levels, whereby traditional statistical methods, such as logistic regression perform very well for credit scoring (see, e.g. Table 1, or [10], for a more comprehensive comparison).

Table 1: Comparison of the performance of different techniques [3].

<table>
<thead>
<tr>
<th>Authors</th>
<th>Lin Reg</th>
<th>Log Reg</th>
<th>RPA</th>
<th>LP</th>
<th>ANN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Henley (1995)</td>
<td>43.4</td>
<td>43.3</td>
<td>43.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Boyle et al. (1992)</td>
<td>77.5</td>
<td>-</td>
<td>75</td>
<td>74.7</td>
<td>-</td>
</tr>
<tr>
<td>Srinivasan &amp; Kim (1987)</td>
<td>87.5</td>
<td>89.3</td>
<td>93.2</td>
<td>86.1</td>
<td>-</td>
</tr>
<tr>
<td>Yobas et al. (1997)</td>
<td>68.4</td>
<td>-</td>
<td>62.3</td>
<td>-</td>
<td>62.0</td>
</tr>
<tr>
<td>Desai et al. (1997)</td>
<td>66.5</td>
<td>67.3</td>
<td>67.3</td>
<td>-</td>
<td>66.4</td>
</tr>
</tbody>
</table>

Hence, considering the relatively sound performance of logistic regression, and the fact that several authors consider logistic regression to be one of the main stalwarts of today’s scorecard builders (see, e.g. [3], [1]), possibly because of its straight-forward character, we have decided to tackle the binary classification problem by means of a logistic regression model. Technically, we can represent logistic regression analysis (also called logit analysis) as a regression technique where the dependent variable is a latent variable, and only a dummy variable \( y_i \) can be observed [11]:

\[
y_i = \begin{cases} 
1 & \text{if the borrower defaults} \\
0 & \text{if the borrower is able to refund his debt}
\end{cases}
\]

The parameters \((b_0, b_1, \ldots, b_k)\) of the \(k\) predictive characteristics used, are then typically estimated using the maximum-likelihood procedure, and the default probability can be expressed as follows:

\[
P(y = 1 \mid X) = P(Y = 1(X) = \frac{1}{1 + e^{-(b_0 + b_1X_1 + \ldots + b_kX_k)}}
\]

### 2.2 Performance measurement

The large majority of credit-scoring studies restricts itself to reporting classification-accuracy measures when evaluating credit-score quality. However, since business objectives do not necessarily match pure classification quality, in this paper, we have attempted to quantify the impact that a modification of the current credit-score algorithm has on business objectives. Hence, we will start by describing classification accuracy, but the understanding will be enriched by mainly graphical interpretations of the modifications on credit risk (i.e. default rate), defaulters’ debt, sales and profit.

Concerning classification accuracy, and consistent with recent studies where performance measures for classification are crucial (see, e.g. [6]), we will use the area under the receiver operating characteristics curve (AUC) to evaluate
predictive performance. Note that, in contrast to the evaluation of the percentage of orders classified correctly (PCC), the receiver operating characteristic curve illustrates the behavior of a classifier without regard to one specific threshold, so it effectively decouples classification performance from this factor (see e.g. [12] for more details, or [1] for an overview of related performance assessment tools in the credit-scoring domain). An intuitive interpretation of the AUC is that it provides an estimate of the probability that a randomly chosen defaulter is correctly rated (i.e. ranked) higher than a randomly selected non-defaulter. Thus, the performance measure is calculated on the total ranking instead of a discrete version of it, so it is clearly independent of any threshold applied ex-post. Note that this probability equals 0.5 when a random ranking is used.

Concerning an evaluation on business objectives, since the company database allowed to trace back all revenues and costs to the individual orders, we were able to determine the exact profitability at the sales-order level. In this paper, we define (default) risk as the percentage of orders that is not profitable within a certain group of orders. Since risk and profitability are by definition the critical performance measure in a credit-scoring business context, we included them as the main credit-scoring performance measures used in this study.

2.3 Resampling procedure

Throughout the different empirical analyses of this study, a resampling procedure was used to assess the variance of the performance indicators. Considering a low proportion of defaulters in the data set used, in this study we will draw samples of \( n \) points without replacement from the \( n \) points in the original data set, allocating an equal amount of defaulters to training as to holdout samples. Hence, we will use a stratified resampling procedure, while using 60% of the data for training the model and 40% for validation purposes. This repartitioning of the data will be performed 100 times, and the differences between the different models will be computed within every iteration (i.e. paired comparisons). The total set of resampling performance indicators was used (i) for estimating the true predictive performance, and (ii) for selecting an adequate model that serves for explaining the business objectives (cf. infra).

3 Data description and variable creation

3.1 Data description

For our research, we used data of a large Belgian direct-mail company offering consumer credit to its customers. Its catalogue offers articles in categories as diverse as furniture, electronics, gardening and DIY equipment and jewelry. We performed the modeling at a moment when the former credit score – constructed by an international company specialized in consumer credit scoring – was about to be updated since it had been in use for 6 years.

For modeling purposes, we will use data of all short-term credit orders placed between July 1st 2000 and February 1st 2002, and their credit repayment
information until February 1\textsuperscript{st} 2003. Within this period, all the credits observed had to be refunded, so it was possible to indicate good versus bad credit repayment within 12 months of follow-up.

3.2 Variable creation

Considering the importance involved in correctly predicting credit-default behavior, a large set of possible predictor variables are offered in literature, including demographics such as age, marital and residential status, occupation of the applicant, financial state of the applicant, expressed in terms of home ownership, debts, incomes and expenditures, the length of the relationship between the applicant and the credit supplier, and previous default information, whether stored in company databases or in credit bureau reports (see, e.g. [7]). However, considering the nature of the credit here, the internal records of a company active in the direct-mail industry can be considered ‘richer’ than that of e.g. banking institutions. Hence, for this classification case, we were able to create a set of 150 possible predictor variables on the company’s database, including almost all the abovementioned variables. Additionally, two possibly important variable categories were included into the list. Firstly, we included variables that are able to assess the nature of the customer’s relationship with the company (i.e. variables related to Customer Relationship Management (CRM), such as recency, frequency, monetary value, the customer’s history of returned goods, increases versus decreases in spending behavior and the number of previous credits refunded). Secondly, besides assessing the creditworthiness of the order, we have attempted to capture the risk of the current order, including issues such as the total amount to pay, the number of months before the order will turn profitable and the default percentage of the product categories that were ordered (based on the defaulting history of all orders). Thus, we have created a set of variables that was designed to optimally predict the risk of a given order at the company.

4 Credit-score construction and evaluation

4.1 Credit-score construction

Three different models were constructed in order to measure the improvement of a modification of the scoring system. As a benchmark and first model, the parameters of the previous score were simply applied to the former set of (20) variables. In a second model, the parameters of these variables were re-estimated on the current data, in order to measure the loss of accuracy due to population drift. Indeed, the company’s managers were detecting a deficiency of the former credit-scoring mechanism, and this could have been due to outdated parameter estimates. Finally, in a third model, a new credit score was proposed, using as a dependent variable whether the order was profitable or not 12 months after the order had been accepted. However, since some orders may contribute more than others to the bottom line, all observations in the training sample were weighed
using the absolute value of the profit they generated. Hence, we included the degree to which the order impacted the result of the company as a weight into the logit model. In order to test the significance of the difference between predictive (AUC) performance, we have applied the resampling procedure as described above, and we selected the model with the mean performance accuracy to be investigated into detail in terms of business objectives.

4.2 Credit-score evaluation

4.2.1 Predictive accuracy
In terms of predictive accuracy, the new model clearly outperforms the previous credit score, as can be seen in Table 2. While the AUC performance of the models is reported on the left side of the table, the differences between the models are shown on the right. Note that these differences are average differences across 100 resamples. They are statistically significant at a significance level of <0.0001.

<table>
<thead>
<tr>
<th>Previous Credit Score</th>
<th>Updated Parameters</th>
<th>New Credit Score</th>
<th>Population Drift</th>
<th>Model Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>.7466</td>
<td>.7791</td>
<td>.8441</td>
<td>.0325</td>
<td>.0650</td>
</tr>
</tbody>
</table>

Note that also the ROC curves were inspected, additionally providing evidence that, on the level of predictive accuracy, the new credit score clearly outperforms the previous model, and the update of the previous model. Hence, we can conclude that, due to the creation of the new predictive variables, the new credit score offers an improved capability of discriminating between 'good' and 'bad' credit applications, and that the improvements seem to be reasonably large, especially when compared to the improvements reported when evaluating an alternative scoring technique (see Table 1).

4.2.2 Business objectives
As mentioned before, previous research often lacks the impact of a modification of the credit score on business objectives. Since the company database provided us with sufficient information to calculate the profitability at the order level, here we have the opportunity to do so. Since credit companies are frequently evaluated based on their total default rates and their debt on defaulted orders, Figure 1 presents the impact that the new score has on these two concepts related to credit risk. In these graphs, the orders are ranked according to their risk, representing the orders that should be accepted due to a low-risk score on the left, and the orders that should be rejected on the basis of a high-risk score on the right. Both are cumulative graphs, meaning that the percent of defaulters if all applicants would be accepted rises to 1.52 percent, while the value lost due to these defaulted orders would mount to 155,548 €. Hence, it should be clear that also here, the new score proves to be performing better than the previous model. For example, if only 5% of the most risky orders (according to the new model)
would have been rejected, the default percentage on the total population would drop by one third of its value, and up to 45% of the value lost due to defaulting would have been avoided. In order to obtain the same effect using the previous score, about 15 to 20% of the most risky orders should have been rejected.

In order to obtain the same effect using the previous score, about 15 to 20% of the most risky orders should have been rejected.

Figure 1: Impact of the models on credit risk.

Hence, these graphs clearly support the findings concerning predictive accuracy. However, in a direct-mail context, rejecting credit applicants also has an impact on profit through lost sales, considering a number of refused orders that would have been repaid flawlessly. This impact is shown in Figure 2.

Figure 2: Impact of the models on revenues and profit.

Surprisingly, the modification of the score seems to have an opposite effect on predictive performance than on company profitability. While the new scoring system is better able to avoid losses due to bad payments, the score tends to be more cautious towards accepting high-value orders – despite the fact that orders contribute to the score proportional to their profitability. Consequently, rejecting high-risk orders leads to an important decrease in sales when compared with the decrease when using the previous scoring system. And since in this specific situation, the revenues stemming from rejected orders are far more important than the loss resulting from defaulting credits, the influence is persistent in the evaluation of company profitability.

While this might lead the reader to conclude that an update of the credit evaluation system will have a negative impact on company performance, this is decidedly not the case. So far, we have evaluated profit for a given acceptance
rate. However, since most credit-scoring institutions use credit risk as the main lever for credit decisions, the above figures can be combined into a joint graphical representation. Figure 3 provides the reader with the evidence of this reasoning.

Suppose the company decides to lower the default percentage on their customer database from 1.52 to 1 percent. In the case of the previous model, this would mean that 85.76 percent of the orders would be accepted for credit, resulting in a total profit of 2,020,437 €, while using the new model would lead to accepting 95.82 % of the credit applicants, representing a total profit of 2,232,270 €. Thus, for a given risk rate, profits would be slightly over 10 % higher when using the new credit-scoring system. By using linear interpolation, we present an overview table for the profit differences that would result when different risk levels are set by management.

Table 3: Joint evaluation of credit risk and profitability.

<table>
<thead>
<tr>
<th>Accepted Risk</th>
<th>Acceptance Rate Previous Model</th>
<th>Acceptance Rate New Model</th>
<th>Percentage Profit Increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.014</td>
<td>0.9564</td>
<td>0.9903</td>
<td>0.0363</td>
</tr>
<tr>
<td>0.012</td>
<td>0.8991</td>
<td>0.9742</td>
<td>0.0865</td>
</tr>
<tr>
<td>0.010</td>
<td>0.8576</td>
<td>0.9582</td>
<td>0.1048</td>
</tr>
<tr>
<td>0.008</td>
<td>0.8210</td>
<td>0.9502</td>
<td>0.1659</td>
</tr>
<tr>
<td>0.006</td>
<td>0.6564</td>
<td>0.8400</td>
<td>0.2309</td>
</tr>
<tr>
<td>0.004</td>
<td>0.4842</td>
<td>0.6996</td>
<td>0.3411</td>
</tr>
</tbody>
</table>
5 Conclusions

In this paper, we have provided evidence against the common practice to evaluate credit-scoring performance based on predictive performance only. Since the acceptance of credit applications has a drastic impact on the cost side – through the debt of defaulted orders – as well as the revenue side – through the sales forgone from rejected applicants – no credit score should be implemented without a thorough understanding of the impact it will have on the company’s risk level, revenues and profits. Furthermore, we have proven that, even when an updated model clearly outperforms the previous model in terms of classification accuracy, it is still the business objective that should determine whether the new scoring system should be applied.

Nevertheless, some arguments can be made for the trade-off between risk and profitability. Both issues are crucial to the success of credit-scoring institutions, and it seems at least conceptually attractive for a company active in credit scoring to opt for the model that delivers the highest profitability for a given default percentage. In this reasoning, the acceptance rate, and the cutoff value that should be applied to enforce the accept/reject decision can be considered as merely the output of a managerial decision consisting of the desired risk rate, and offers the necessary link for understanding the interaction between risk and profitability.

6 Issues for further research

It seems at least intriguing that exactly the model that was designed to detect profitable versus non-profitable orders shows a lower profit compared with the other models for a given acceptance rate. Hence, it seems striking that impact of the score on the company’s operations depends more on the business objective than on mere classifier performance. It is also intriguing that a model that is built on the profit of an individual order does not clearly outperform a non-profit directed model on an aggregated profit level. All these issues definitely seem to suggest that any decision concerning credit-scoring performance can hardly be based on a mere technical ranking performance indicator, yet should optimally be taken only when all implications of the new score have been investigated. Further research might be directed towards improving the credit-scoring algorithm to include the different outcomes during the training process.

References


