Applications of Data Mining in E-Business and Finance
C. Soares et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-890-8-45

A Tripartite Scorecard for the Pay/No pay Decision-Making in the Retail Banking Industry

Maria Rocha SOUSA [a] and Joaquim Pinto da COSTA [a]
[a] *Faculdade Ciências Universidade Porto, Porto, Portugal*
e-mail: maria.rochasousa@portugalmail.pt
e-mail: jpcosta@fc.up.pt

Abstract. Traditionally retail banks have supported the credit decision-making on scorecards developed for predicting default in a six-month period or more. However, the underlying pay/no pay cycles justify a decision in a 30-day period. In this work several classification models are built on this assumption. We start by assessing binary scorecards, assigning credit applicants to good or bad risk classes according to their record of defaulting. The detection of a critical region between good and bad risk classes, together with the opportunity of manually classifying some of the credit applicants, led us to develop a tripartite scorecard, with a third output class, the review class, in-between the good and bad classes. With this model 87% decisions are automated, which compares favourably with the 79% automation rate of the actual scorecards.

Keywords. Pay/no pay decision, mass-market, tripartite scorecard

Introduction

The ubiquity of digital communications has led to the generalization of online payments in individuals' Demand Deposit Accounts (DDAs). Retail banks have to assure a prompt answer for those payment requests, which can be in the order of millions a day. When the DDA has insufficient balance the bank has to decide whether or not to pay that debit transaction (a pay/no pay decision-making) in a process named Non Sufficient Funds (NSF). This pay/no pay decision must be performed at the latest by the end of the day, to fit the Financial Net Settlement System service level's requirements. Optimizing this decision-making entails the decision to be uniform, objective and fast, with the minimum of mistakes and losses.

Currently at a retail bank, most of the decisions (79%) are automatically managed, while critical decisions are left for manual assessment. However, the automatic behavioural scoring models in use were developed to predict default in a six-month period; furthermore, to keep the implementation straightforward, they do not entirely emulate human reasoning. Therefore, some distinctive features of the problem are not materialized in them. Both customers' earnings and NSF process cycles take one month to be completed. Hence, if a "pay" decision is made, it is expected that the DDA cures within

30 days (the DDA is cured when it does not exceed its balance and overdraft limits). This led us to consider the development of a specific model to classify short-term credit risk for mass-market customers of the retail bank.

1. A Credit Model for the Pay/no pay Decision

The research summarized here was conducted by using a large real-life dataset (comprising 187733 records) from the credit data of the leading Portuguese retail bank. We choose the SAS Enterprise Miner package to perform all computations. Each customer's DDA in the sample was labelled as *good* if it cured within 30 days and as *bad*, otherwise. Rejected transactions were also included in the sample. They were assigned to *good* or *bad* classes according to their balance in the following month. The adopted class definition is based on Portuguese economic practices, as well as specific market segments, and therefore to the pattern of NSF and customers' earnings cycles.

In pay/no pay decision-making, human evaluation is usually supported in the existing information of the three-month period preceding the decision day. This human procedure was incorporated in the models using a three-month observation window. The sample comprises DDAs with pay/no pay decisions of an entire month, the decision period. For those DDAs, data were collected for the previous three-month period – the observation window. The performance was evaluated for each customer's DDA according to his behaviour in the 30-day period after having had a pay/no pay decision – the performance window [1]. The driving idea was to look to pay/no pay decisions in the decision period and evaluate whether the DDA cured in the following 30-day period. If so, the DDA was labelled as non-defaulter; if not, as defaulter.

The information gathered for each costumer's DDA comprises 47 characteristics related with the DDA transactional pattern (e.g. the structure and volume of monthly debits and credits and balance cycles) and customers' behaviour in their relation with credit. The current or past flawed experiences with financial institutions were included in the sample as well, such as missing payments and bankruptcy status.

The original sample dataset was randomly divided into three subsets: 70% for the training set, 20% and 10% to be the validation and test groups, respectively. The proportion of each target class in the actual population, 18% defaulter and 82% non-defaulter, was kept in the sample dataset.

The classifiers were trained both with an equal loss matrix and a loss matrix that integrates the cost of misclassification, empirically estimated using a sample of historical decisions.

1.1. Estimation of the loss matrix

In this application models are required to minimize the total loss associated with the decisions, rather than the number of errors. One of the most efficient approaches to build models that are sensitive to non-uniform costs of errors is to make the decision based on both the estimated probabilities of the unseen instances and a measure of business performance (profit, loss, volume of acquisitions, etc) [2]. We adopted the expected loss value for each possible decision.

For each approval in pay/no pay decision-making, the bank charges an amount M that equals the maximum between a fixed fee and the interests. When the transaction is a cheque, the bank charges an additional fee: f_+ if the cheque is paid, f_- otherwise. The estimation of the loss matrix was based on the following principles:[1]

- The error of classifying an actual defaulter as non-defaulter generates a loss that is equal to the value of the transaction;[2] since the mean value of cheques is higher, the costs of misclassifications was differentiated by group of transactions. Therefore, the expected cost of a bad decision in cheques, l_c, and the expected cost of a bad decision in other cases, l_o, was weighted by the expected proportions of cheques and other transactions in the true population, p_c and p_o. The expected loss is therefore $p_c\, l_c + (1 - p_c) l_o$.
- The error of classifying an actual non-defaulter as defaulter produces a loss corresponding to the fees that the Bank dos not charge/collect and the revenue from charging the fee f_-, in the case of cheque refusal. Weighting those fees by the corresponding proportion, the loss is given by $p_c\,(f_+ - f_-) + M$.

Although fees and interests are pre-defined, some scenarios can correspond to exclusions, decreasing the amount to be charged. Hence, rather than using the standard predefined fees, which would lead to unrealistic and inflated profits, matrix parameters were estimated empirically using a sample of historical decisions. Mean charged fees and the expected costs were then calculated for each of the two groups, cheques and others. Loss matrix parameters (normalized values) were estimated as 0.28, 0.22, 0.44, and 0.34 for p_c, M, f_+ and f_-, respectively. The normalized cost of misclassification is 18 if the transaction is a cheque, and is 10 for other type of transactions.

These principles allow a practical evaluation of the expected loss of a single decision in the pay/no pay decision-making, and can be summarized in a loss matrix that puts more weight on costumers wrongly predicted as non-defaulters with the proportion 1:49.

2. Binary Scorecard

Several standard binary classification models, based on logistic regression, classification trees [3] and neural networks [4,5], were designed from the same input dataset. More than just discriminating between the two classes, the models yielded a scored dataset as a result of their training.[3] Two different strategies were gauged: training the models to estimate only the probabilities of each class of the target variable, without incorporating any business objectives for which the predictor will be used. This strategy corresponds in adopting the equal-loss matrix, with which both types of errors are equally weighted. In a second strategy the training of the models incorporates the estimated business costs,

[1] Our approach for evaluating the loss of a pay/no pay decision does not incorporate indirect profits such as commercial benefits from keeping relation with good customers active, neither the costs of preserving bad customers. Although quantifying them would lead to valuable results it would also require considering some non-trivial business assumptions. As that would take us beyond the objectives of the current work, they were not considered.

[2] Although in practice the loss of misclassifying a defaulter is less than the value of the transaction, we considered the worst scenario in which the credit is totally lost.

[3] A scored dataset consists of a set of posterior probabilities for each level of the target variable, corresponding to the probability of defaulting and not defaulting.

focusing not in the minimization of the misclassification rate but in the optimization of the profit or loss. The selection of the cutoff for each case is easily determined. If the probability of defaulting p_d of a given costumer is known, the best cutoff for a general loss matrix $\begin{bmatrix} l_1 & l_2 \\ l_3 & l_4 \end{bmatrix}$ is determined by comparing the expected loss of predicting as defaulter, $l_1\ p_d + l_3\ (1 - p_d)$, with the expected loss of predicting as non-defaulter, $l_2\ p_d + l_4\ (1 - p_d)$. The resulting cutoff is $\left(1 + \frac{l_2 - l_1}{l_3 - l_4}\right)^{-1}$. For the equal-loss matrix case, the threshold is 0.5; for the estimated loss matrix, the threshold is 0.02. The best results are summarized in Table 1. For each model, the minimum loss, the sensitivity (percentage of actual defaulters predicted as defaulters), the specificity (percentage of actual non-defaulters predicted as non-defaulters), and the error rate are provided. As a reference performance, the results for two baseline classifiers are also presented. The Naive 1 model refuses all examples, while Naïve 2 model classifies all as non-defaulters.

Table 1. Results for the best binary models.

(a) Minimization of business rules.

Model	Loss	Specificity	Sensitivity	Error rate
Logistic Regression	0.724	27.2%	98.6%	59.9%
Decision Tree	0.697	28.4%	98.8%	58.9%
Neural Network	0.697	34.3%	98.2%	54.1%
Naïve 1	0.820	0.0%	100.0%	82.0%
Naïve 2	8.820	100.0%	0.0%	18.0%

(b) Minimization of the error rate.

Model	Loss	Specificity	Sensitivity	Error rate
Logistic Regression	0.094	98.4%	55.2%	9.4%
Decision Tree	0.083	97.8%	64.5%	8.3%
Neural Network	0.089	98.1%	59.6%	8.9%
Naive 1	0.820	0.0%	100.0%	82.0%
Naive 2	0.180	100.0%	0.0%	18.0%

Models tuned with the matrix incorporating the business rules have high sensitivity (above 98%), while their specificity is low (below 35%). This strategy led to models with high error rates. When the models were developed to minimize the error rate the results were essentially reversed. The error rate of these models is mostly due to misclassified defaulters in the set.

2.1. Tripartite Scorecard

The results attained with the binary classifiers show that none could discriminate the defaulter from the non-defaulter in a satisfactory way. We also observed a certain overlap between the distribution of the defaulter and of the non-defaulter, when analysed over the predicted probability of defaulting, meaning that the models were not effective in distinguishing them. When varying the cutoff value we are just trading off between the two types of possible errors. Pushing the cutoff near the values obtained for the estimated matrix, almost all defaulters are correctly predicted, while most of the non-defaulters are incorrectly predicted as defaulters. Relaxing the cutoff to values around the value obtained for the default matrix, the errors are reversed. When deploying a system of this kind in a retail bank, there is the opportunity of defining a third type of decision, the review class: an example predicted as review will be evaluated manually by human experts, possibly making use of some additional information. Therefore, we investigated the possibility of designing models with three output classes [6]: defaulter, review and non-defaulter. Bipartite and tripartite scorecards have been used in the industry before, but only in an ad hoc way, with no effort being made to find the optimal division [7].

Table 2. Confusion matrix for a three-output model.

	Predicted Defaulter	Review	Predicted Non-Defaulter
True Defaulter	p_1	p_2	p_3
True Non-Defaulter	p_4	p_5	p_6

Considering the generic confusion matrix for a three-output model (Table 2) the training of the models was driven to find two cutoffs simultaneously that provide low error rates p_3 and p_4 (assuming that all manual decisions are correct) and high automation rate. The lack of standard formulations and implementations to solve a problem of this kind, led us to start with a simple approach. Starting on the models previously designed, the two cutoffs were determined as follows:

- A cutoff was initialized as 0.0. Next, it was iteratively raised until a predefined probability p_3 (= 0.025) was obtained.
- A cutoff was initialized as 1.0. Next, it was iteratively lowered until a predefined probability of error p_4 (= 0.050) was obtained.

Finally, the percentage of automatic correct decisions ($p_1 + p_6$), the percentage of defaulters in the approved set (p_3/p_6), the error rate ($p_3 + p_4$) and the automation ($p_1 + p_3 + p_4 + p_6$) were measured. Three models, presented in Table 3, were chosen from all under evaluation.

Table 3. Tripartite Scorecard Results.

	Cutoff	Cutoff	Specificity	Sensitivity	Approved	Error	Automation
Model	Low	High			defaulters	rate	
Logistic Regression	10.5%	30.5%	92.6%	84.8%	3.5%	7.2%	82.1%
Decision Tree	7.0%	23.7%	93.2%	80.2%	4.9%	8.0%	86.6%
Neural Network	11.0%	27.3%	92.8%	84.2%	3.7%	7.4%	85.0%

The three-class output models have more balanced measures of sensitivity and specificity, with a better prediction of the true classes. About 15% of the decisions, corresponding to the overlapping region, are left for human assessment.

Assuming that the percentage of actual defaulters approved automatically does not consider the effects of the recovery actions that can be performed, we accepted a value up to 5%. The Tree based model was considered the most adequate for the pay/no pay decision-making, providing 87% of automatic decisions. Furthermore, a Decision Tree model is suitable for deployment and explanation of the decisions.

3. Discussion

This study focuses on the development of a scorecard for supporting the evaluation of default risk in the pay/no pay decision-making of a retail bank.

Binary classification models were developed based on well-known classification techniques. Although an extensive study was conducted, the attained discrimination between the two classes (default and non-default) was not satisfactory. When the weights of the two types of errors are heavily asymmetric, the boundary between the two classes is pushed near values where the highest cost error seldom happens. For equal misclassification losses, the boundary is biased to predict accurately the dominant class. Therefore, the research continued with the development of tripartite scorecards, with a third output class, the review class, in-between the good and bad classes. The final model enables 87% of automatic decisions, comparing favourably with the actual scorecards.

More principled approaches for optimally determining the boundaries between the good, review and bad classes are currently being investigated. A complementary model is also being developed for managing the resulting credit in arrears.

References

[1] Lyn C. Thomas, David B. Edelman, and Jonathan N. Crook. *Credit Scoring and its applications*. SIAM, 2002.

[2] Roger M. Stein. The relationship between default prediction and lending profits: Integrating roc analysis and loan pricing. *Journal of Banking & Finance*, 29:1213–1236, 2005.

[3] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[4] B.D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

[5] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[6] K. B. Schebesch and R. Stecking. Support vector machines for classifying and describing credit applicants: detecting typical and critical regions. *Journal of the Operational Research Society*, 56:1082–1088, 2005.

[7] David J. Hand, So Young Sohn, and Yoonseong Kim. Optimal bipartite scorecards. *Expert Systems with Applications*, 29:684–690, 2005.