Model of credit loss and its use in spread decisions

[(1)]João Fernando Serrajordia Rocha de Mello, MSc
[(2)]Carlos Alberto de Bragança Pereira, Phd
University of São Paulo, Brazil

Analytical methods for granting credit have gone through great advances in recent decades, particularly in the field of statistical methods for classification of individuals into groups with different default rate. Most of the existing works suggest decisions of the type granting credit or not, regard just marginally the expected financial outcome of the operation.

This work aims to propose a model for credit risk evaluation that takes into account a broader output than the traditional Credit Scoring ones. It provides a more detailed and longitudinal view about the future performance of a credit agreement, which goes beyond the classification of default or non-default. In addition to this improvement, it also aims to expand the problem decision domain, leaving a binary response (accept/reject the claim) to respond to the following question: "What is the just rate to cover a given risk"?

First the methodology classifies a credit operation in one of three groups according to how the contracts are ended: finished agreement by paying as scheduled, with renegotiation efforts and by write-off. Afterwards, it provides an expectation of ending time for the contract. The model is built in such a way that it provides model parameter estimates by the same computational algorithm that solves logistic regression. Some further analysis is proposed, such as calculating the expected result of a portfolio, determining credit proposals that are accepted/rejected and determining the minimum interest rate for the contract acceptance. These analyses are based on assumptions about alternative financial outcomes of the transaction, speculating its ending time and the way it is finished.

Keywords: Credit Scoring, Survival Analysis, Decision Theory, Longitudinal Data, Multinomial Modelling

(1) Institute of Mathematics and Statistics – University of São Paulo
(2) Institute of Mathematics and Statistics – University of São Paulo

1. Introduction

Usually credit applicants are classified into two groups: Bad and Not Bad. Such classification is commonly based on historical credit contracts. The old contracts establish what a bad and a not bad profile is. These profile classifications are based on statistical models that use several variables whose values could be obtained at the credit application time. For instance, the logistic regression is the most well known technique for this purpose. See, for instance, Thomas, Edelman & Crook (2004).

Considering all historical portfolios, some financial institutions calculate the averages of losses and of profits, respectively, for the bad and not bad clients. A compromise between profits and losses defines the rule for which the credit decision is taken.

The objective of this paper is to present a method for predicting outcomes of new contracts. The novelty of our method is to consider the stepwise probabilities of the instalment payments. In addition the method takes into account the way the contracts were ended: loss, renegotiation or without default.

With the features described above, one can use the method together with some decision theory aspects, to calculate the *optimal* spread for a specific contract/profile.

Some backgrounds are presented in Section 2. Section 3 describes the multinomial classification into three performing groups, the stepwise instalment payment probabilities estimation and the calculation of the expected outcome. With simulated data, based in real observed situations, Section 4 presents the application of the method developed in the former sections. Section 5 is devoted to some interesting considerations for future research.

2. Background

This session presents the basis of the proposed methodology: basic assumptions, formulas for result calculations, notations, and data structure. The reader can use this session as both: a prologue for the following sessions that contains the main methodology or as a reference guide while reading the methodology independently.

1.1.Outcome calculation

In the case of a paid loan with no occurrence of default or aging on payments, its outcome can easily be calculated from the price table (or the amortization system used), its operational costs can be deduced, and the financial funding can be taken into account and so on. But that seldom occurs. However, many events can occur during the loan payment time.

We assume that any loan can be classified into one of three groups, and for each group its outcome can be calculated (or approximated) given the number of instalments paid when the loan is ended and the basic information of the contract (spread, term, etc).

The three groups are defined according to how the loan is ended. From now on, we call them Performing Groups, and they are defined as follows:

Group 1 – No occurrence of default;
Group 2 – Occurrence of default and need of credit recovery efforts;
Group 3 – Loans classified as Write Off and considered loss.

1.2.Notation

The following notation is used to explain the model:

i. The *i* index (positive integer) is used to indicate the loans in the training sample;
ii. The *j* index (also an integer) is used indicate the Performing Group of the loans;
iii. The *k* index (also an integer) is used to indicate the instalments of the loans;
iv. The *t* is an auxiliary index used when 2 indexes are needed for the instalments;
v. The *l* index (also an integer) is used to indicate the covariates;
vi. The vector of ***p*** observed covariates is denoted by $\mathbf{x_i} = (x_{i1};x_{i2};\ldots;x_{ip})$. We have here the i^{th} individual that belongs to the *n* sized training sample, and $x_{i1},x_{i2},\ldots,x_{ip}$ are the observed values for the covariates used in the model.
vii. The probabilities of the events $Y_i=(1;0;0)$, $Y_i=(0;1;0)$ and $Y_i=(0;0;1)$ are the probability that the ***i*th** loan belongs to Group 1, 2 or 3, respectively. Hence, the probability that the loan *i* finishes belonging to Group j (j=1,2,3) is denoted by the simplified notation $\pi_{ij} = P(Y_{ij} =1)$. To the i^{th} unit there is an associated vector of probabilities, $\mathbf{\Pi_i} = (\pi_{i1};\pi_{i2};\pi_{i3})$, that sum one; $\pi_{i1}+\pi_{i2}+\pi_{i3} =1$. $\mathbf{Y_i}$ is then a multivariate Bernoulli random vector with order 3 and dimension 2. Since the loans are

considered independent observations of this Bernoulli random vector its sum over all units, the vector N = $(N_1;N_2;N_3)$ – here N_j is the total number of loans in Group j – is a trinomial with parameters $n = N_1+N_2+N_3$, the total number of loans (sample size) in the study, and $\mathbf{\Pi_i}$, the vector of probabilities;

viii. G_j is the set of N_j loans in the sample that performed in Group j;

ix. A new parameterization simplifies the algebraic calculus due to the strong properties of multinomial distributions. Let the conditional probability of the unit i belong to Group 2 given that it does not belong to Group 3 be represented by the following notation: $\theta_{i2} = P(Y_{i2}=1|Y_{i3}=0) = \pi_{i2} \div (1-\pi_{i3})$. Writing now the original parameter we have $\mathbf{\Pi_i} = (\pi_{i1};\pi_{i2};\pi_{i3}) = ([1-\pi_{i3}][1-\theta_{i2}];[1-\pi_{i3}]\theta_{i2};\pi_{i3})$;

x. The number of instalments of loan i is denoted by S_i. S_i =12 could indicate that the contract of the i^{th} loan has of one year of monthly payments;

xi. If the loan i ends in either Performing Group 1 or 2, the number of instalments paid at the end of the loan contract is represented by T_i for the random variable and t_i for the observed value. If the Performing Group is 3, then T_i is the number of instalments paid plus 1. As in Groups 1 and 2 the number of instalments paid at the end of the loan vary from 1 to S_i, and in Group 3 it varies from 0 to S_i-1, this keeps the coherence of our notation, in order that T_i always vary from 1 to S_i.

xii. $\mathbf{T}$ is the vector of instalment information of the sample: $\mathbf{T}=(T_1;T_2;...;T_n)$.

xiii. Z_{ik} is a function of T_i. It is equal to 0 if $1 \leq k < T_i$ and to 1 if $k=T_i$. Z_{ik} is not defined for $k > T_i$;

xiv. O_i is the financial outcome of the loan i.

1.3.Data structure

In order to avoid data confidentiality and have them freely available, the data was generated simulating a real situation according to the experience of the authors.

We simulated a set of n accepted loan applications made over 12 consecutive months. All these loans are instalment types and the customers are from a legal entity retail portfolio. The loans are paid monthly in instalments, and the terms are either 6, 12, 18 or 24 months.

The covariate vectors $\mathbf{x_1,...,x_n}$ were generated by a special pair of functions in order to have a correlation structure. The variables observations were generated according to a

random model similar to the proposed in this paper, but with nuisance parameters to avoid an artificial fit. The full algorithm is described by Mello (2009).

3. Methodology

The basic tool of this article is the logistic regression that is shortly introduced below. It is explained how to use it as a multinomial classification model in order to estimate the probabilities that a contract belongs to each of the three Performing Groups. Another use of the logistic regression is the evaluation of the probabilities that a contract finishes having each possible number of instalments paid. Subsequently, we describe the simple method used to define the optimal spread for each loan that is based on the estimated probabilities.

2.1.Logistic Regression

The logistic regression (Hosmer & Lemeshow, 1989) is the most commonly used statistical model for predicting binary responses in credit scoring. It is used to evaluate the probability (π_i) of an event, usually default, according to a vector of covariates $\mathbf{x_i}$. It is a particular case of the Generalized Linear Models, the link function being the logit, resulting in the following function:

$$ln\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x_i'\beta}$$

Here, $\boldsymbol{\beta}$ is a vector of parameters associated to each one of the covariates. The estimates of $\boldsymbol{\beta}$ are obtained by maximum likelihood with a training sample of n observations $\mathbf{x_i}$, i=1,…n, of a binary response variable. From this equation we obtain the following expression:

$$\pi_i = \frac{1}{1+e^{-\mathbf{x_i'\beta}}}$$

2.2.Multinomial Classification Model

Now an expansion of the well known logistic regression model as used in the traditional credit scoring systems is presented. Instead of classifying loans as bad or not bad accounts, the aim is to classify them into one of the three Groups defined above.

The technique consists of developing a logistic model to estimate π_{i3}, the probability of loan i performing as Group 3. The second step is to consider only the sample with the

other two groups and perform another logistic regression to estimate θ_{i2}. Recall that θ_{i2} is the probability that the i^{th} loan performs as Group 2, given that it does not belong to Group 3. Since $(\pi_{i1};\pi_{i2};\pi_{i3}) = ([1-\pi_{i3}][1-\theta_{i2}];[1-\pi_{i3}]\theta_{i2};\pi_{i3})$, we know that

$$P(Y_{i3}=1) = \pi_{i3} \text{ and } P(Y_{i2}=1|Y_{i3}=0) = \theta_{i2}.$$

Consequently, having a sample with *n* loans, the likelihood function of the parameter vector $(\pi_{i1};\pi_{i2};\pi_{i3})$ can be written as function of π_{i3} and θ_{i2}. It can be shown that this likelihood function can be factored into two parts: one depending only on π_{i3} and the other depending only on θ_{i2} – see Pereira & Stern (2008) and Mello (2009). Having these two parameters in independent varying fields, they can be estimated independently.

Now, writing π_{i3} and θ_{i2} as logistical models, we have:

$$ln\left(\frac{\pi_{i3}}{1-\pi_{i3}}\right) = \alpha_1 + X_{i1}\beta_{11} + \cdots + X_{in}\beta_{1n} \qquad (3.1)$$

$$ln\left(\frac{\theta_{i2}}{1-\theta_{i2}}\right) = \alpha_2 + X_{i1}\beta_{21} + \cdots + X_{in}\beta_{2n} \qquad (3.2)$$

The parameters $(\alpha_1;\beta_{11};\ldots;\beta_{1p})$ and $(\alpha_2;\beta_{21};\ldots;\beta_{2p})$ can be estimated separately in their logistic regression models. Our estimates of π_{i3} and θ_{i2} are based on the maximum likelihood estimators obtained by these two logistic regressions as follows:

$$\hat{\pi}_{i3} = \frac{1}{1+e^{-(\hat{\alpha}_1+X_{i1}\hat{\beta}_{11}+\cdots+X_{in}\hat{\beta}_{1n})}} \qquad (3.3)$$

$$\hat{\theta}_{i2} = \frac{1}{1+e^{-(\hat{\alpha}_2+X_{i1}\hat{\beta}_{21}+\cdots+X_{in}\hat{\beta}_{2n})}} \qquad (3.4)$$

Hence, $\hat{\boldsymbol{\Pi}}_i = (\hat{\pi}_{i1},;\hat{\pi}_{i2};\hat{\pi}_{i3}) = \left([1-\hat{\pi}_{i3}]\left[1-\hat{\theta}_{i2}\right];[1-\hat{\pi}_{i3}]\hat{\theta}_{i2};\hat{\pi}_{i3}\right)$

Parameter interpretation:

$e^{\beta_{l1}}$ is the odds ratio for the events Y_{i3}=1 due to the increase of the covariate X_l by one, keeping all the other constant.

$e^{\beta_{l2}}$ is the odds ratio for the event Y_{i2}=1 given Y_{i3}=0, due to the increase of the covariate X_l by 1, keeping all the other variables constant.

α_1 is the intercept of the model for π_{i3}: equal to $\ln(\pi_{i3}/(1-\pi_{i3}))$ for $X_{i1}=...=X_{ip}= 0$.

α_2 is the intercept of the model for θ_{i2}: equal to $\ln(\theta_{i2}/(1-\theta_{i2}))$ for $X_{i1} =...=X_{ip} = 0$.

2.3.Model for the Number of Instalments Paid at the End of the Loan

To be more accurate on the prediction of the future outcome of a new loan, it is also necessary to estimate the probability of the loan *i* ending by the k^{th} instalment (*k* varying from 0 to the total numbers of instalments of the loan) according to its profile given by its values for the components of $\mathbf{x_i}$.

Since the three groups have different ending reasons for their loans, we are going to develop three models by conditioning to each Performing Group.

This model is very similar to the proportional hazards model, Cox & Oakes (1984) and Hosmer & Lemeshow (1999). Our model is based in the number of instalments paid at the end of the loan instead of the time and, of course, this is a discrete variable.

Given the Performing Group *j*, we define the hazard of the loan ending by the instalment *k* as:

$$h_{ijk} = P\left(T_i = k \middle| T_i > k-1; \boldsymbol{x_i}; Y_{ij} = 1\right) \tag{3.5}$$

Remember that Y_{ij}=1 means that the loan ***i*** performs in the $\boldsymbol{j^{th}}$ group.

Given that a loan is in Group *j*, the probability of the loan finishing before paying the instalment *t* (*t* =1,...,S_i) evaluated by the credit application is as follows:

$$P\left(T_i = t \middle| \boldsymbol{x_i}; Y_{ij} = 1\right) = h_{ijt} \prod_{k=1}^{t-1} \left[1 - h_{ijk}\right] \tag{3.6}$$

Let *T* be the maximum term of all loans in the sample (T=Max(S_i), i=1,... n). We define n indicating variables:

$$D_{itk} \text{indicating observation on instalment } k, D_{itk} = \begin{cases} 1 \text{ if } t = k \\ 0 \text{ otherwise} \end{cases}$$

In order to estimate h_{ijk} according to the different profiles given by the covariates $\mathbf{x_i}$, we can state a logistic model for h_{ijk}:α

$$ln\left(\frac{h_{ijk}}{1-h_{ijk}}\right) = D_{i1k}\alpha'_{j1} + \cdots + D_{iTk}\alpha'_{jT} + x_{i1}\beta'_{j1} + \cdots + x_{ip}\beta'_{jp} \quad (3.7)$$

Or in a more straightforward equation,

$$h_{ijk} = \left\{1 + exp\left[-\left(D_{i1k}\alpha'_{j1} + \cdots + D_{iTk}\alpha'_{jT} + x_{i1}\beta'_{j1} + \cdots + x_{ip}\beta'_{jp}\right)\right]\right\}^{-1} \quad (3.8)$$

Where all the *α's* and *β's* in (3.7) and (3.8) are parameters of the model, given the Performing Group *j*. Hence, we have three separate linear models; one for each *j* (=1, 2, 3).

Using maximum likelihood estimators, $\hat{\alpha}'s$ and $\hat{\beta}'s$, the following equation gives the maximum likelihood estimates of the rates h_{ijk}:

$$\hat{h}_{ijk} = \left\{1 + exp\left[-\left(D_{i1k}\hat{\alpha}'_{j1} + \cdots + D_{iTk}\hat{\alpha}'_{jT} + x_{i1}\hat{\beta}'_{j1} + \cdots + x_{ip}\hat{\beta}'_{jp}\right)\right]\right\}^{-1} \quad (3.9)$$

Parameter interpretation:

The parameters $\boldsymbol{\alpha_{j1},\ldots,\alpha_{jT}}$ work like *T* intercepts and they give the baseline logit hazard rate over the *T* possible instalments, given the Performing Group *j*. They have an analogy with the baseline hazard function of the Cox's proportional hazards model.

The parameters $\boldsymbol{\beta_{ij}}$ are the variation rates on the logit hazard due to the increase of x_{iL} by 1 (*L=1,...,p*), given the Performing Group *j*.

Parameter estimation

Given the Performing Group *j*, Maximum likelihood estimations for $\boldsymbol{\beta'_j}$ can be obtained by maximizing the likelihood:

$$L\left(\boldsymbol{T}; \boldsymbol{\beta'_j} | \boldsymbol{X}; Y_{ij} = 1\right) = \prod_{i \in G_j} P\left(T_i = t_i | \boldsymbol{X}; \boldsymbol{\beta'_j}; Y_{ij} = 1\right)$$

$$= \prod_{i \in G_j} \left[h_{ijt_i} \prod_{k=1}^{t_i - 1} \left(1 - h_{ijk}\right) \right] \quad (3.10)$$

This likelihood can be written using an instalment view. In such a way, each new row will be an instalment. As we have defined in Session 2, the variables Z_{ik} = 1 if T_i=*k* and 0

otherwise. Note that, for each loan *i*, the variables Z_{ik} are observed just up to the first *one* is obtained when the observational process stops for that loan.

Given the performing Group *j* (*j* =1, 2 or 3), we can re-write the likelihood function as:

$$L\left(\boldsymbol{T};\boldsymbol{\beta_j}'|\boldsymbol{X};Y_{ij}=1\right)=\prod_{i=1}^{n}\left[\prod_{k=1}^{t_i}{h_{ijk}}^{Z_{ik}}\left(1-h_{ijk}\right)^{1-Z_{ik}}\right] \quad (3.11)$$

This likelihood is the same as the logistic regression likelihood, which means that we can estimate their parameters with the same algorithms as the common logistic regression, needing just to have a previous preparation on our data set.

A similar model was introduced by Singer and Willet (1993).

Data preparation

Each row of our simulated dataset represents a loan. The data preparation consists of repeating each row T_i times, so that each new row will represent an instalment. We also need to create the *T* dummy variables D_{itk} with *k* varying between the replied rows of the loan (*k*=1,...t_i), and *t* varying within each row, assuming values between 1 and T. In addition, it is necessary to build the answer variables Z_{ik} as defined above.

Then, for each loan *i*, we will have t_i answer variables, one for each instalment until the loan is finished. Tables 3.1 and 3.2 exemplify that data preparation. Table 3.1 is the original data table with one row for each loan and Table 3.2 is the prepared table with several rows for each loan, one for each instalment.

Table 2.1. Original Data Table Example

Loan(i)	Term (t)	t	Y_1
1	12	1	1
2	6	3	1
3	24	2	1

Table 2.2. Prepared Data Table Example

Loan(i)	Term	Instalment (t)	Z	D_1	D_2	D_3	...	D_{24}
1	12	1	1	1	0	0	...	0
2	6	1	0	1	0	0	...	0
2	6	2	0	0	1	0	...	0
2	6	3	1	0	0	1	...	0
3	24	1	0	1	0	0	...	0
3	24	2	1	0	1	0	...	0

Observe that on Table 3.2 the loan 1 appears once because it was finished by the first instalment. Also, there are three rows for loan 2 and two rows for loan 3 because they ended by the third and second instalments respectively.

Expected outcome

This methodology is based on the assumption that we can calculate the expected outcome as a function of the number of instalments paid by the end of the loan and the performing group in which it is finished. That is:

$$E(O_i|T_i, = t_i\ Y_{i1} = 1, \mathbf{x_i}) = S(i|\mathbf{x_i})$$

$$E(O_i|T_i, = t_i\ Y_{i2} = 1, \mathbf{x_i}) = S(t_i|X_i) - C(t_i|\mathbf{x_i})$$

$$E(O_i|T_i, = t_i\ Y_{i3} = 1, \mathbf{x_i}) = S(t_i|\mathbf{x_i}) - C(t_i|\mathbf{x_i}) - B(t_i|\mathbf{x_i})$$

Here, $E\left(O_i\middle|T_i = t_i\ Y_{ij} = 1, \mathbf{x_i}\right)$ is the expected outcome of loan *i,* given that it finished by the t_i^{th} instalment, performing in the Group *j* having covariate vector $\mathbf{x_i}$;

$S(t_i|\mathbf{x_i})$ is the accumulated spread in all instalments paid. Observe that $S(t_i|\mathbf{x_i})$ takes into account the spread of all instalments paid, that is, if the Performing Group is 1 or 2, the instalment paid are the first to the t_i^{th}, and if the Performing Group is the third, the instalments paid are the first to the t_i-1^{th} if t_i>1 and 0 if t_i=1.;

$C(t_i|\mathbf{x_i})$ is the cost due to credit recovery efforts of a contract ended just before paying the t_i^{th} instalment;

$B(t_i|\mathbf{x_i})$ is the principal balance left before paying the instalment t_i.

The quantities $S(t_i|\mathbf{x_i})$ and $B(t_i|\mathbf{x_i})$ can be calculated based on the PRICE table (or the proper amortization system). The quantity $C(t_i|\mathbf{x_i})$ is calculated as a simple linear function over the balance left. This kind of model is not the focus of this work, but one can use a more complex model that fits ones particular business better.

Actually, what we want to calculate $E(O_i|\mathbf{x_i})$ for each instalment. Using conditional probability we have:

$$E(O_i|\mathbf{x_i}) = \sum_{j=1}^{3}\sum_{t=1}^{T_i} E\left(O_i\middle|T_i, = t\,; Y_{ij} = 1; \mathbf{x_i}\right)P\left(T_i, = t\,|Y_{ij} = 1; \mathbf{x_i}\right)P(Y_{ij} = 1|\mathbf{x_i}) \qquad (3.12)$$

Using the estimates described earlier we can estimate the expected outcome of the loan *i* as:

$$\hat{E}(O_i|\mathbf{x_i}) = \sum_{j=1}^{3}\sum_{t=1}^{T_i} E\left(O_i\middle|T_i, = t\,; Y_{ij} = 1; \mathbf{x_i}\right)\hat{P}\left(T_i, = t\,|Y_{ij} = 1, \mathbf{x_i}\right)\hat{P}\left(Y_{ij} = 1|\mathbf{x_i}\right) \qquad (3.13)$$

2.4.Spread definition

The definition of spread is based on the minimum spread the bank may ask for a given operation that gives as outcome the minimum result the bank accepts. In this paper we assume that the aimed outcome is already set. And so, this session discusses how to use the described model to have the minimum interest fee to have the accepted outcome.

As described before, it is assumed that we can calculate the outcome of a contract given the group in which it was finished, the number of instalments already paid and the contractual spread. As we have estimates of the probabilities for the Performing Groups and for any possible number of instalments paid at the end of the loan, we can also calculate the expected outcome given the spread.

Then, we can use a simple iterative method as the dichotomy method to find out what is the minimum interest fee so that the expected outcome equals the aimed outcome.

4. Results

In this section we have the estimated models for the contracts performance: the conditional multinomial classification model and the instalments prediction model. Also

some diagnosis for these models and some further analysis concerning the expected outcome of a set of contracts and the suggested spread are presented.

3.1.Multinomial classification model results

As discussed earlier, this model is based on two logistic regressions, one for prediction of contracts that will perform as belonging to Group 3 and the second for prediction of contracts performing as belonging to Group 2 given that they will not belong to Group 3.

Applying the methodology to these data we have the estimated parameters as shown on the following table:

The linear parameters for the logit of P(Y_{i3}=1) was estimated by maximum likelihood. The Table 4.1 shows the variables considered in the model, the degrees of freedom for the parameter estimation, their respective parameters, their standard error, the Wald statistic and its descriptive level for the null hypothesis of the parameter equals zero.

Table 3.1. Parameter estimates of the model for P(Y_3=1)

Parameter	*DF*	*Estimate*	*Standard error*	*Wald's Statistic*	*p-value*
Term6	1	-2,523	0,159	251,2	<0,0001
Term12	1	-2,265	0,152	223,0	<0,0001
Term18	1	-1,567	0,151	107,4	<0,0001
Term24	1	-0,465	0,147	9,979	0,0020
Positive History	1	-1,334	0,083	260,2	<0,0001
Entity's age	1	-0,084	0,005	292,4	<0,0001
Borrowing	1	-0,457	0,218	4,399	0,0360
Negative History	1	1,565	0,064	598,0	<0,0001

Table 3.2. Parameter estimates of the model for P(Y_2=1| Y_3=0)

Parameter	*DF*	*Estimate*	*Standard Error*	*Wald's statistic*	*p-value*
Term6	1	-1.964	0.1151	291.1	<.0001
Term12	1	-1.604	0.1073	223.6	<.0001
Term18	1	-1.215	0.1088	124.7	<.0001
Term24	1	-0.550	0.1104	24.9	<.0001
Positive History	1	-0.725	0.0468	239.3	<.0001
Entity's age	1	-0.030	0.0040	53.8	<.0001
Age*Neg. Hist.	1	0.001	0.0037	6.9	0.0085
Negative History	1	0.558	0.0870	41.1	<.0001

3.2.Instalments model results

Just as discussed previously, the parameters for this model where estimates with the same algorithm that resolves logistic regression. We have one equation for each of the ending groups as long as the model gives the probabilities for each possible number of instalments paid at the ending of the contract given the ending group.

The estimates are exposed in a similar table as in the last session:

Table 3.3. Estimates for the Instalment Prevision Model for Group 1

Parameter	DF	Estimate	Standard Error	Wald's Statistic	p-value
Intercept	1	-2.2694	0.0982	533.52	<.0001
Term12	1	-0.5943	0.1128	27.77	<.0001
Term18	1	-1.5731	0.1207	169.75	<.0001
Term24	1	-1.972	0.1350	213.26	<.0001
T	1	0.3346	0.0261	164.03	<.0001
T*Term12	1	-0.1114	0.0277	16.18	<.0001
T*Term18	1	-0.0865	0.0269	10.33	0.0013
T*Term24	1	-0.1597	0.0268	35.60	<.0001
Positive History	1	-0.6239	0.0201	962.94	<.0001
Entity's age	1	-0.0495	0.0014	1217.36	<.0001
Borrowing	1	0.3071	0.0720	18.20	<.0001
Negative History	1	0.6478	0.0264	602.69	<.0001

Table 3.4. Estimates for the Instalment Prevision Model for Group 2

Parameter	DF	Estimate	Standard Error	Wald's Statistic	p-value
Intercept	1	-5,1982	0,1950	710,60	<,0001
Term24	1	0,5366	0,2567	4,37	0,0366
T	1	1,1301	0,0485	543,36	<,0001
T*Term12	1	-0,5695	0,0345	271,72	<,0001
T*Term18	1	-0,8012	0,0384	434,34	<,0001
T*Term24	1	-0,9479	0,0497	363,73	<,0001
Positive History	1	-0,6096	0,0576	111,89	<,0001
Entity's age	1	-0,0134	0,0034	15,32	<,0001
Borrowing	1	0,6192	0,1923	10,37	0,0013
Negative history	1	-0,4499	0,0575	61,13	<,0001

Table 3.5. Estimates for the Instalment Prevision Model for Group 3

Parameter	DF	Estimate	Standard Error	Wald's Statistic	p-value
Intercept	1	-1,7662	0,1431	152,2415	<,0001
Term6	1	0,2772	0,1386	4,0007	0,0455
T	1	-0,0682	0,0142	23,0465	<,0001
Borrowing	1	1,2768	0,2417	27,9029	<,0001
Positive History	1	0,5611	0,0656	73,2183	<,0001

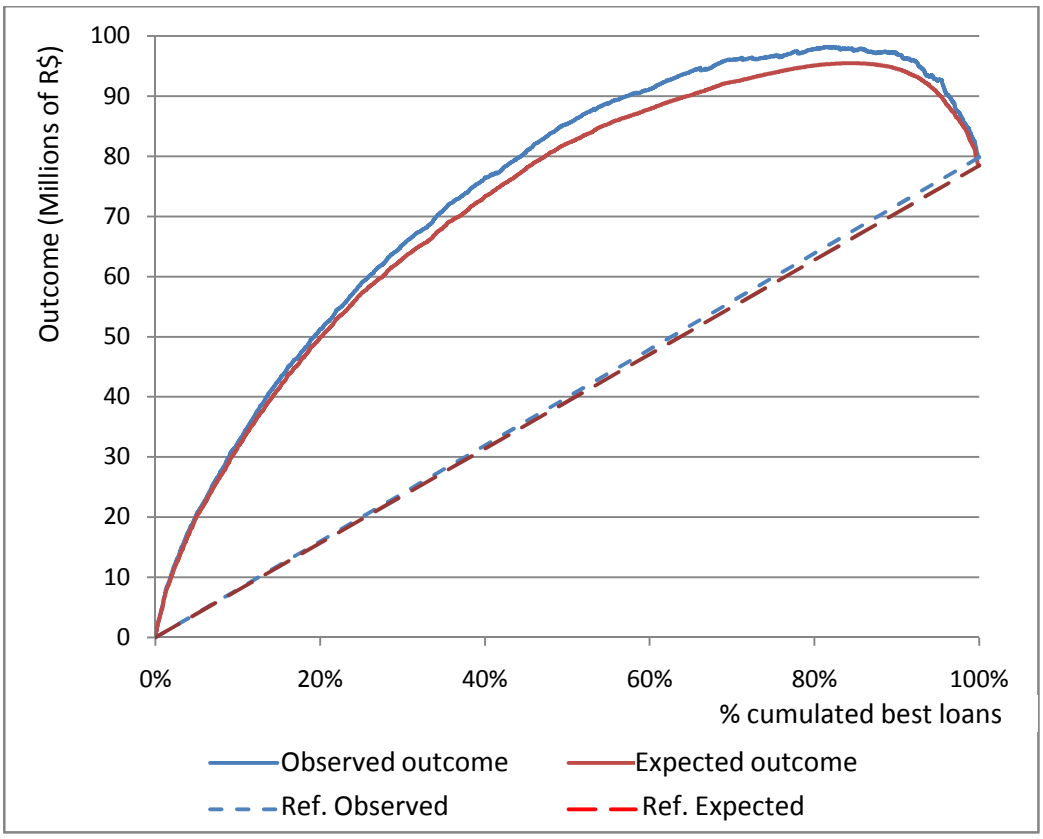
3.3.Expected Outcome Analysis

A simple analysis can be performed through the viewpoint of the Decision Theory, using the utility function as the identity function on the outcome itself. The first analysis assumes that the spread is fixed and the decision space constitutes of the set of customers that are accepted.

Since the method produces expected results, let us take the result over the value borrowed, just to put all the contracts in a comparable scale. Now we can rank the customers by that measure from the worst to the best.

A good way of assessing the quality of a model is the Lift Chart (Vuk & Curk, 2006). We made a little adaptation in this chart for our problem. We rank the accounts by its expected value divided by its initial balance and plot the percentage of the best loans on the abscissa and the cumulative expected result corresponding to these accounts on the ordinate. Also, as a diagnostic chart for the model, we can plot in the same graph a curve for the observed result and check if it gives a good fit. The result is the following plot:

Graph 4.1. Observed and expected lift charts



Observing Graph 4.1, the maximum expected result would be archived granting credit for a little more than the best 80% of the applicants, and it would have had an expected outcome of about 95 million of *Reais*. On the other hand, granting credit for the whole set of customers gave us an expected outcome of almost 80 million of *Reais*.

Also, the expected curve is very near the observed curve indicating a good fit for the ultimate use of the model that concerns the outcome itself.

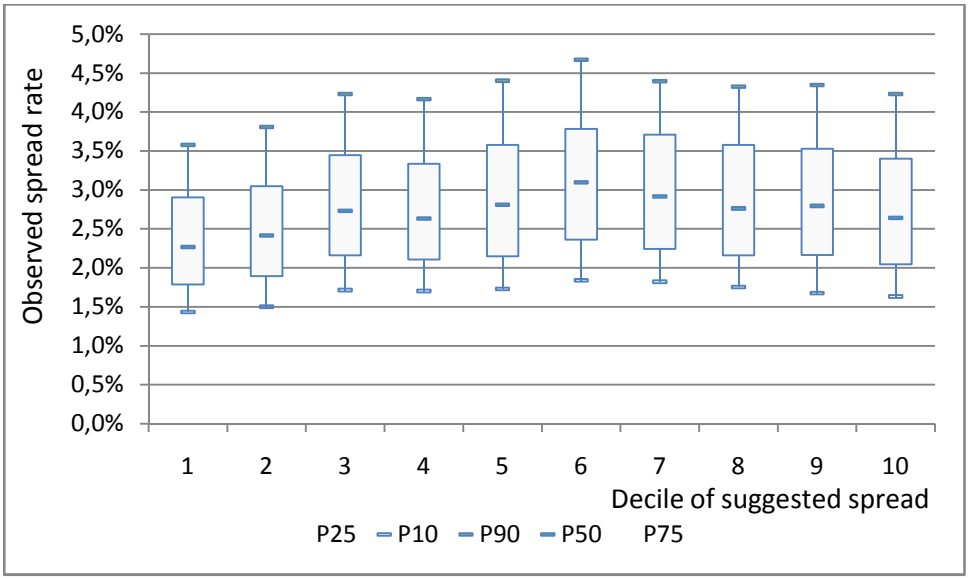
5. Suggested tax analysis

Finally, we would like to consider the decision space of our problem as the tax the bank would need to ask. It is a complex issue the question of how the required spread interferes with the risk. It is a two way relationship: the bank asks for higher spreads because the customer is a high risk one and the higher the spread, the harder the contract is to be paid and the riskier the contract is expected to be. Because of this two way dependency, it is hard to evaluate interesting ways: how does the spread interfere with the risk? Leaving this discussion for another work, we will take as assumption that there is no dependency to build the following analysis.

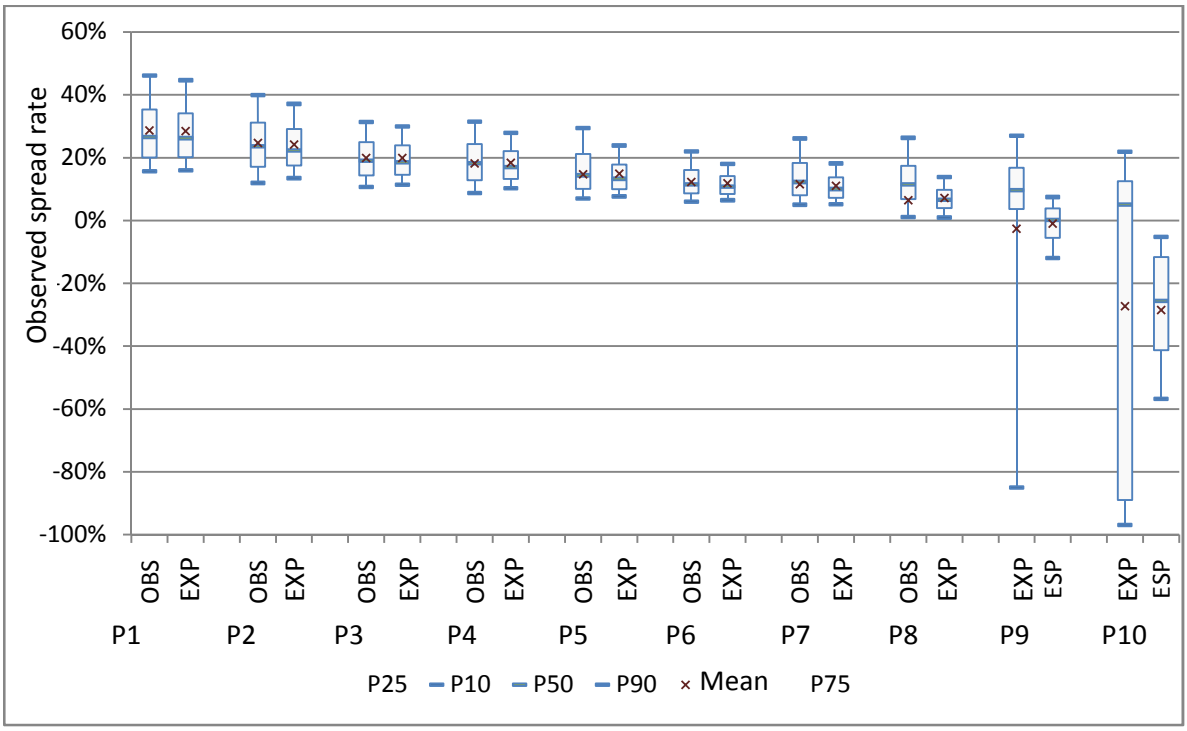
From the financial viewpoint, the greater the spread is the better it seems to be. On the other hand, the less likely the customer is of taking that loan. Consequently, obtaining the minimum acceptable spread gives the sales force a tool for negotiation: the spread would be as great as possible, and it can be lowered to increase the odds for acceptance, but never will be less than the minimum set by the model.

We can split the loan portfolio used for model training into ten groups according to the suggested spread's deciles so that the groups have approximately the same number of loans. Then we can plot, for each group, a *boxplot* with the realized spread to check if the spread decisions made in the past was taken someway according to the credit risk.

Graph 5.1. Suggested versus observed spread



Graph 5.2. Expected result versus suggested spread decile



In the Graph 5.1 we observe that there was no relationship between the realized spread and the suggested spread, evidencing that the decision of spread was made without taking into account the credit risk.

Using the presented group of percentiles of suggested spread, we can then build a boxplot of the expected result and, if available for diagnosis, and another for the observed result. The resulting plot is illustrated by Graph 5.2:

We observe in the Graph 5.2 still a good fit of the proposed model as long as the observed result is similar to the expected result. In addition we observe the two higher spread groups with negative expected results suggesting that these contracts would not be accepted, at least not with these taxes.

6. Final Considerations

The methodology developed here seems to work appropriately for loans made in instalments, however some credit institution will ask for additional sophistication that takes into account recovery costs, discounts and so on.

The present model did not consider accounts that, at the time of data collection, were still running and could also give some historical occurrences. This problem can be overcome by taking multinomial models with missing data as presented by Paulino and Pereira (1995) and was mentioned by Mello (2009).

It is still an open issue the way to include relationships between spread and credit risk. It may be interesting to work with reasonable conditions and build scenarios. Necessarily, those set of conditions should include the one discussed here: the case of zero correlation.

References

Cox DR & Oakes D (1984), ***Analysis of Survival Data,*** **London:** Chapman & Hall.

Hosmer DW & Lemeshow JS (1999), ***Applied survival analysis: regression modelling of time to event data****,* NY: Willey.

Hosmer DW & Lemeshow J (1989), ***Applied logistic regression,*** NY: Willey.

Mello JF (2009), **Predictive Model for credit loss and its application to spread decisions**, MSc Thesis, S Paulo, Brazil: Instit Math & Stat - U São Paulo (portuguese).

Paulino CD & Pereira CA (1995), Bayesian methods for categorical data under informative general censoring, ***Biometrika*** 82(2) 439-46.

Pereira CAB & Stern JM (2008), Special characterizations of standard discrete distributions, ***REVSTAT - Statistics J*** 6:199-230.

Singer JD & Willet JB (1993), It´s About Time: Using Discrete-Time Survival Analysis to Study Duration and Timming of Events, ***J Educational Statistics*** 18(2):155-95.

Thomas LC Edelman DB & Crook JN (2004), ***Readings in Credit Scoring: Foundations, Developments, and Aims****.* Oxford: Oxford University Press.

Vuk M & Curk T (2006), ROC Curve, Lift Chart and Calibration Plot, ***Metodoloski zvezki*** 3(1):89-108.