UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

# MÉTODOS PARA ESTIMAR RIESGO CREDITICIO EN BASE A MINERÍA DE DATOS Y TEORÍA DE JUEGOS

**TESIS POR COMPENDIO DE PUBLICACIONES PARA OPTAR AL GRADO DE DOCTOR EN SISTEMAS DE INGENIERÍA**

CRISTIÁN DANILO BRAVO ROMÁN

PROFESOR GUIA:
RICHARD WEBER HAAS

MIEMBROS DE LA COMISIÓN:
LYN C. THOMAS
JOSÉ MIGUEL CRUZ GONZÁLEZ
ALEJANDRO JOFRÉ CÁCERES
BART BAESENS

Este trabajo ha sido financiado por Beca Doctorado Nacional CONICYT Nro. 21090573

SANTIAGO DE CHILE
OCTUBRE 2012

To my engine, and my life.

# Acknowledgements

A PhD is certainly a major investment in the life of any person. It represents making several sacrifices, personal and financial, for the goal of an enormous reward: the intellectual one. After four years taking a PhD I guess the greatest realization I made was how little I know. I entered the PhD thinking I knew all there was to know about data mining, and marketing applications; I looked the PhD as an specialization on one topic that was attractive to me (credit scoring), and for learning a bit more about techniques that I already "mastered". I was seriously wrong. Four years after deciding to take a PhD, I realize the sea of knowledge in which we swim, and I also realize the impossibility of any one person or group to really master any particular broad field. We are forced to choose a (very) narrow strip of knowledge, in hope that after dedicating a life to it we are able to understand a relevant number of its complexities, and maybe are lucky enough to transmit this knowledge to others. That is the responsibility of the scholar, and that is the life I was actually choosing when taking the academic path, and I hope I can become, many decades from now, a researcher such as the ones I have been able to meet during my very short career.

I would like to acknowledge some of the people that have accompanied, mentored, or supported me during the development of this thesis, in many different ways. To prof. José Miguel Cruz, for the trust deposited in me, and the professional tutoring, reminding me that our kind of science is not far from practical applications, and also challenging me to give all that I can. To professor Lyn C. Thomas, for being a real mentor, a wise man who is capable of transmitting his wisdom in a very simple, kind, and direct manner, and possibly the most brilliant person I have ever had the luck of crossing paths with. I am honoured to have studied with you, and I hope my work reflects this. And finally, to my advisor professor Richard Weber, for what can only be summarized as teaching me how to be a researcher, in all its dimensions. I have learned that being a researcher includes many different sides: creating knowledge, transmitting it, connecting with practice, forming new profes-

sionals, getting funding, creating connections worldwide, having fun, go nights without sleeping, working hard, and a much, much more. All this is thanks to him, Vielen Dank, Sie sind ein Freund!

In a personal fashion I would like to thank three persons who have been with me all this time. First, my mother, for transmitting every value I have, which sums so much into very few words. To my grandfather, for showing me the importance of family, and how that is the most important thing in one person's life. And finally, to Cindy, for sharing with me happiness unbounded, and forging a life project with me, you are the engine in my life. ¡Los amo de verdad!

Finally, some persons I cannot forget to include: Sebastián, a true partner in crime during all these years, and we have so many fun things to do forward! To Julie Lagos, for being the invisible hand that makes everything work. To my siblings, for being always with me. To the Finance Center, for accepting me and giving me a group to which I am proud to belong to. And to everyone else I might have left out, you know who you are. You've been great, I've been data minin', thank you, and good night!

# Resumen Ejecutivo

Medir la probabilidad de no pago de un gran número de solicitantes de crédito, el llamado riesgo de crédito a consumidores, es un problema clásico de la gestión financiera. Este problema requiere de una gran cantidad de herramientas estadísticas que lo hacen idóneo para su estudio por el área de *Business Analytics*. Su análisis se justifica en el fuerte impacto que los créditos a consumidores tienen en el mercado, pues por ejemplo en Chile más del 50% de los créditos se encuentran en carteras masivas, y en el resto del mundo se estima que superan a los créditos comerciales en más de un 50%. Esta tesis estudia este problema en base a la formalización de lo que se conoce sobre las determinantes del no pago (*default*) y la transformación de ese conocimiento en herramientas estadísticas para la medición del riesgo.

Para lo anterior, durante el doctorado desarrollé una sucesión de publicaciones con el fin de unir el modelamiento económico con la práctica estadística predictiva, formalizando el conocido hecho que existen no pagadores por razones de falta de capacidad de pago, y por problemas en voluntad de pago. El trabajo comenzó generando un modelo económico que captura las utilidades de los solicitantes y la entidad prestamista, utilizando esta nueva definición de no pagadores, resultando en una serie de restricciones que definen un espacio de créditos factibles. Luego, los no pagadores son asignados a estas dos clases utilizando un nuevo método de *clustering* semi-supervisado que los agrupa tanto por sus características estadísticas como por su comportamiento económico, reflejado a través de las restricciones generadas previamente.

El fin último de esta separación es mejorar la clasificación de los créditos y la comprensión del default, por lo que el siguiente paso correspondió al estudio de métodos de clasificación con múltiples clases. Para ello se exploró inicialmente la regresión logística multinomial, y luego se profundizó en el análisis de supervivencia, estudiando teóricamente los métodos de riesgos en competencia y los modelos mixtos, y desarrollando herramientas computacionales liberadas públicamente para apoyar futuras aplicaciones.

Los métodos propuestos mejoran entre 1%-10% la discriminación por sobre los métodos clásicos en bases de datos reales, y enriquecen fuertemente la comprensión del default a través de las nuevas variables significativas y los patrones encontrados. Más aún, los modelos y herramientas desarrollados pueden ser perfectamente extrapolados a otras disciplinas, pues este trabajo ha mostrado cómo se puede enriquecer la clasificación donde típicamente se cuenta con dos clases, por la vía de añadir conocimiento adicional acerca de comportamientos económicos observados.

# Executive Summary

Measuring the probability that a large number of borrowers return a given loan, commonly referred to as consumer credit scoring, is a classical financial management problem that is attractive to Business Analytics, given the large number of statistical tools that are required to solve it. The study of consumer credit risk is justified by the large impact that personal loans have in the economy: in Chile, they represent more than 50% of all granted loans, and it has been estimated that worldwide they exceed corporate loans by 50%. This thesis focuses on this problem, formalizing knowledge regarding default and transforming that knowledge into statistical tools for risk measurement.

During my PhD I have published a series of studies that join economical modelling with predictive statistics, formalizing the well-known fact that default is caused by problems in capacity of repayment, or by problems in willingness to repay. The study starts by generating an economical model that captures the utilities of the borrowers and the lender, under this new definition of default, resulting in a set of constraints that define a space of feasible loans. Next, defaulters are segmented using the constraints previously generated as the input for a new semi-supervised clustering algorithm, grouping them into two classes by both statistical properties and repayment behaviour.

The final goals of this segmentation are to improve the classification accuracy of borrowers, and to enhance the comprehension of default, so the next step was to study multi-class classification methods. Initially, multinomial logistic regression was studied, followed by a deeper study of survival analysis that researched the theoretical properties of competing risks and mixture models. Publicly available computational tools were developed for those techniques, in order to support future applications in this or other fields.

The proposed methods present a classification accuracy improvement of between 1%-10% over classical techniques in tested datasets, and the analysis of obtained significant variables and patterns found strongly enrich default comprehension. Furthermore, the developed models and tools can be extrapolated to other disciplines, since this work has shown a general methodology to enrich classification when typically only two classes are present, by adding additional knowledge regarding observed economical behaviours.

# Contents

# List of Figures

# Chapter 1

# Introduction

Credit scoring corresponds to the use of statistical models to transform relevant data into numerical measures that guide credit decisions (Anderson, 2007), and its main objective is to estimate the probability of default, i.e. the event of a customer not paying back a granted loan in a given time period. Credit scoring is of paramount importance in any financial market, since it streamlines the process of granting loans, enabling discrimination based on quantitative analysis rather than on subjective criteria. Applying these systems gives consumers a fair chance at accessing loans, and also gives a quantitative tool to lenders for controlling losses, thus increasing efficiency and facilitating greater competition, which makes a direct impact on society.

There are several branches of credit scoring, since different tools and models are needed to measure different sources of credit risk for every financial instrument, and research on this topic is as wide and varied as there are financial instruments. This work is focused on Consumer Credit Risk, i.e., when a lender grants loans to a relatively large pool of consumers, so that each individual loan has a very small effect on the full loan portfolio. In order to control risk in this setting, it is necessary to apply statistical methods and data mining tools, making credit scoring one of the "classical" and best known applications of data mining.

A credit scoring system for new customers is commonly referred to as "application scoring", whereas a system built for customers for whom information is already known is referred to as "behavioural scoring". This thesis focuses on improving application scoring, that is, improving credit

scoring aimed at borrowers for whom very little information is known, since there is no available credit history. These systems are commonly based on socio-economic conditions, public debt databases if available – in Chile the database of the Superintendency of Banks and Financial Institutions (SBIF) keeps such records –, and also other (not very popular) databases such as the public arrears database that Equifax (DICOM) maintains. Until now, the discussion has been focused on the variables that better determine consumer credit risk, but there is only so much information that can be extracted from these databases, creating opportunities for more sophisticated approaches.

The work reflected in this thesis focuses on applying state-of-the-art data mining techniques to credit scoring in an original fashion to improve existing results. This is done through the development of a combined data mining/game theory model, attempting to include more information about the dynamics and behaviours of the customers and the lender, and observing this behaviour in operational databases. The research was made available to the community through a series of publications which comprise this thesis.

This chapter introduces the problem to be studied. In particular, a general introduction to data mining in the context of credit scoring is presented in section 1.1, followed by the objectives in the next section. The resulting publications will be summarized in section 1.3, and the last section presents a full list of the academic contributions that were developed during the author's period of study for the PhD degree.

## 1.1 Consumer Credit Scoring and Knowledge Discovery in Databases[1]

To construct a credit scoring system, it is common to follow the Knowledge Discovery in Databases (KDD) methodology (Fayyad et al., 1996). The process, shown in Figure 1.1, is a general guide that facilitates obtaining knowledge from large quantities of unstructured data. The KDD is commonly considered to be an instruction manual of the model building process, but this is not the case. It is more a train of thought than a true manual, guiding the modeller in asking him or herself the correct questions.

[1]This section is based on the paper "Granting and Managing Loans for Micro-Entrepreneurs", currently in its

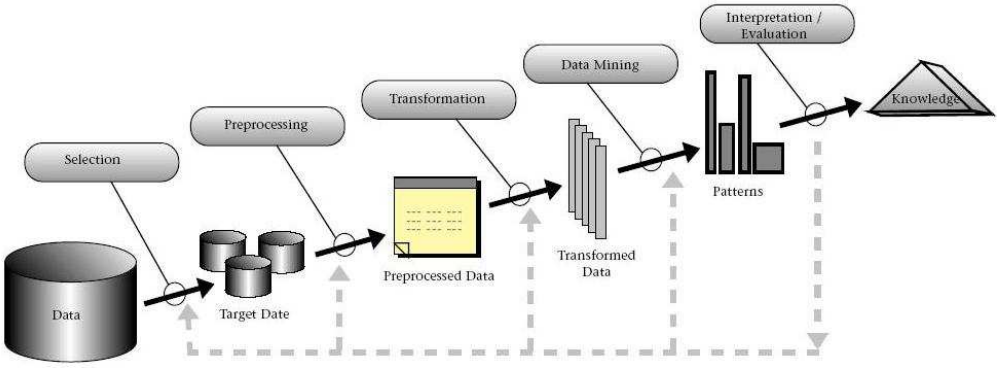



Figure 1.1: Knowledge Discovery in Databases (KDD) process.

The process starts from rough, commonly unstructured, data spread among several databases. From them, a successive, and sometimes iterative, process of cleaning, filtering, and transforming data converges into a clean *tableau*, or data matrix, from which models can be derived. The process of creating statistical models from such a matrix is what is known as data mining. After the data mining process, the models have to be validated and implemented. The following subsections describe these steps in detail, and incorporate the particularities of credit scoring in the context of KDD.

### 1.1.1 Problem Definition

Mackinon and Glick (1999) recommend that after identifying the problem to be studied (i.e. default), the general goal that has to be accomplished must be defined. For example, a typical credit scoring goal might be "to obtain a global decrease of default rates for the mass consumer portfolios". This goal must then be divided into specific objectives, and quantifiable measures must be defined, such as "default rates should fall no less than 5%, contrasted with current rates".

When the objectives have been set, a quantifiable objective variable should be defined, especially in the case of predictive modelling. The objective variable is the quantitative measure of the studied event, and has to be calculated in the tableau for each case. A good objective variable has to satisfy the following:

second revision at the European Journal of Operational Research. The full paper can be seen in Appendix A.

- It must be *homogeneous*: All cases should measure the same event. For example, in consumer credit scoring the most widely-used objective variable is modelling default as failure to repay an instalment for more than 90 days after the due date. In marketing, it might be buying a certain product or product category.
- It must be *unbiased*: All cases should have the same time frames, and should possess the same chances for the studied event to occur. This means that if a portfolio of loans has different terms (12 months, 24 months, and so on), a common horizon of observation must be defined for all of them. The typical notion is to consider default only if it occurs 12 months after the granting of the loan, or during the first twelve instalments if the loan is repaid monthly. In marketing, the periods might be shorter, considering purchases by trimester, or by month.

### 1.1.2 Data Set Consolidation

After the problem has been defined, the next step is to identify the relevant databases in which the available information is scattered, and then extract all relevant variables, to finally develop the repository which is used for constructing the models. Data should be selected from all available sources, be they external (such as market data, for-sell databases, or other), internal databases (such as in-house operational databases, or data warehouses), and variables designed and created from those sources specifically for the problem to be studied, typically in the form of ratios, averages, or other consolidation operations. The main goal of this step is to ensure variance and credibility at the same time: the variables must describe the elements studied, and the source must be reliable so that the results it may give can be trusted. The last point is particularly important, since including an incorrect variable in this step might lead to incorrectly specified models farther on, and that misspecification will probably not be apparent to the modeller.

The input in this phase corresponds then to every available source of data, while the output is a database with all available information for solving the problem. The data should be consolidated on a per case basis, per loan in the case of credit scoring, thus forming the tableau.

The final condition for developing an effective credit scoring system is that a homogeneous and representative sample of the population for both classes (defaulters and non-defaulters) is obtained. Based on the relevant literature, e.g. Thomas et al. (2002) or Anderson (2007), it is recommended

to first segment customers and the requested loans by differentiating between new customers in one group and renewing or current customers in the other, resulting in the need for creating both application scoring and behavioural scoring systems.

### 1.1.3 Preprocessing, Data Cleansing, and Variable Selection

With the tableau ready, the next step involves eliminating variables that do not have an impact on the problem at hand, whether due to having little variance, too many null values, or for other reasons. This step is usually the longest in most data mining projects, involving up to 80 percent of the total development time (Myatt and Johnson, 2009). The preprocessing phase of this step commonly includes the following steps:

- Data Description: All different attributes (variables) to be used must be examined and described by using common statistical tools (means, averages, etc.). The intention of this step is to understand and evaluate data quality, since the existence of too many null variables, variables that are too concentrated, or variables with too little variance will be evident.
- Data Cleansing: Each variable should be explored to standardize some entries and eliminate data that should not be part of the attribute. Examples typically include diverse forms of recording values, since for a variable such as “Country”, potential values might include “United Kingdom” and “UK”; or “Chile” and “Chili”. Other inconsistencies might include negative ages, or exaggerated values for income, or out of range values such as high income borrowers.
- Additional Information Inclusion: Sometimes corporate databases might have some known inconsistencies that have to be taken into account, or some attributes that might need to be consolidated. This step of the process is the best time to do so.

The tableau at this point is clean of errors, complete, and ready to be analysed using statistical tools. The next step consists of data preparation and variable (feature) selection. For this, a three-step procedure was developed through several projects during the period of study for the PhD: minimizing the risk of eliminating potentially useful variables selecting features in a cascade-like

approach, and maximizing the knowledge extracted from the respective dataset. Descriptions of the three steps follow:

1. Analysis of Concentration of Features and Missing Values: In order to discard useless variables quickly, those concentrated in a single value in more than 99% of the cases, and those missing more than 30% of their values were eliminated. The rationale of the second criterion is to reduce the number of discarded cases or imputations realized in order to construct the final model.

2. Univariate analysis: The remaining variables were tested to find distribution equality across groups, using the objective variables (defaulter/non-defaulter) as the splitting criterion. The idea is to discard variables with no apparent capacity to discriminate between the classes, assured by statistical tests. In particular Kolmogorov-Smirnov (K-S) and $\chi^2$-tests were applied to continuous and discrete variables, respectively.

3. Multivariate analysis: In order to analyse multivariate discriminatory capacity, the remaining variables were used as input for a classification tree (Safavian and Landgreve, 1991), again taking default behaviour as the objective variable. This classification tree was allowed to be overfitted, since the goal was to determine any possible multivariate correlation across the dataset. Variables that were not included at any level of the tree were discarded, by the consideration that there was no apparent gain in including them in a final model.

This procedure is valid only when the analysis is supervised, that is, when an actual objective variable is present. For unsupervised models (no objective variable, only exploratory), then only the first step is necessary.

### 1.1.4 Data Transformation

Most statistical models have very strict requirements regarding the types of data they accept. Most algorithms work only on numerical inputs, requiring the dataset to be a subset of $R^V$, with $V$ the number of variables, or even $[0,1]^V$. This step in the process ensures that data is in the correct format. Some common transformations include:

- Normalization: Continuous attributes, such as income or age, might be scaled to a common interval, thus giving every variable the same relative weight in the model. Some methods correspond to normalization to interval $[0,1]$, such as z-scaling – subtracting the mean value of a variable from each value and dividing the result by the standard deviation –, and decimal scaling, which corresponds to divide each variable by $10^m$, with $m$ the number of significant digits the maximum value of the variable has.

- Ordinal variable mapping: Categorical variables with intrinsic ordering, such as a high/-middle/low choice on a survey, can be transformed to numerical values defining an explicit magnitude for each category, representing a conceptual distance by using a numerical one. The risk of following this strategy is that the definition and interpretation of this numerical distance is subjective, and thus can lead to including incorrect patterns in the model.

- Transformation of Categorical Variables: Categorical variables can be represented mathematically based on transforming an attribute with $N$ different categories to $N-1$ binary (dummy) variables, so that if a particular element has category $n$, then variable $n$ takes the value of one and all the rest take zero. Another popular transformation in credit scoring is the Weight of Evidence (WOE, Siddiqi (2006)), which calculates the information value of each category (logarithm of proportion between good and bad payers) and assigns that value to the variable, without dividing it into multiple dummy ones.

- Discretization of Continuous Variables: Some models require that all variables take only discreet values, with the legal requirements of credit scoring being a noteworthy example, so there is a need to transform continuous values to discreet ones. Other reasons for wanting to discretize variables correspond to variables concentrated in two or more segments, or variables with non-linear behaviour in a linear model. Discretization is usually performed by defining intervals of the variable as categories first, and then applying one of the strategies for categorical variables.

- Aggregation: Some variables might increase their discriminative or explicative capacity when they are aggregated with others. Examples might include total cost as the aggregation of variable and fixed costs, or the same with income.

Once this step is finished, a tableau with optimal conditions is available for applying data mining techniques. This tableau optimally contains all records that are desired for modelling, with a

set of potentially explanatory variables that are scaled, transformed, and aggregated/discretized as necessary.

### 1.1.5 Models and Data Mining

The next step is finally to apply statistical models to the tableau. This step is usually an iterative process, consisting of trying several configurations of models, parameters, and variables in combination. As was mentioned before, there are two main types of learning processes: supervised learning, in which the objective is to create a function that allows predicting the class of new cases using data from known ones, and unsupervised learning, where the objective is to discover structures or patterns in the data itself, without any particular event in mind. There is also a new emergent application, semi-supervised learning, that uses additional information to enrich unsupervised data. This thesis includes a novel semi-supervised algorithm, explained in detail in Chapter 3.

In unsupervised learning, typically all of the dataset is used directly for the chosen models. In supervised learning it is common first to segment the database to validate the obtained results. Two strategies are the most widely-used, the first being holdout validation, in which 20% - 40% of data is set aside to validate the model, using an average of 25% according to Hastie et al. (2009). The second method, more thorough but more costly in computational time, is cross-validation, in which the dataset is split into $N$ different pieces, and there are $N$ subsequent training rounds, in each $N-1$ forming the training set (from which the model is derived) and the one left out is used for testing the results. The set used for testing is changed each round, until all have been used. The results are then averaged and returned. Finally, some models require that the training set is further divided to extract a validation set, which is typically used for parameter tuning.

With this division in mind, the process of selecting a model can be divided into:

- Final Attribute Selection: Even though attributes selected in previous steps cannot be discarded as being completely independent of the studied event, not all of them grant a significant increase in a multivariate model, making it necessary to perform model dependent feature selection. Several methods are available for this purpose, with the more common

being backward and forward feature selection (Hosmer and Lemeshow, 2000) which starts with all (or conversely none of) the variables and removes them (conversely adds them) one by one until no improvement occurs, based on significance or accuracy tests.

- Parameter Tuning and Model Selection: The techniques must be selected according to the difficulty of the problem, the complexity required, computing time available, and knowledge of the modeller, among many other variables. Even when a set of models to test is selected, each model may have multiple parameters to tune, so a number of prior experiments might be needed for optimal results. When this is the case, it is common to leave 20% of the dataset out to obtain the final model before conducting full training.

Once the final model is selected, the results can be reported using the test dataset. A model is validated when the test results are consistent, the variance of parameters is robust (using cross validation), and the variables obtained are reasonable in a business sense. The output of this step is a completely validated model that has to be implemented. The next section shows the model most commonly used in the particular case of credit scoring.

#### 1.1.5.1 Logistic Regression in Credit Scoring

The final decision regarding the acceptance of the loan application can be made by comparing the calculated probability of default, estimated using a suitable model and the variables obtained from previous steps, with a suitable pre-defined threshold (Hand and Henley, 1997). Several studies in credit scoring have focused on comparing the classification performance of different techniques (see, for example, Baesens et al. (2003), Finlay (2011)). Their main conclusion is that logistic regression reaches performance levels comparable with more sophisticated data mining approaches for modelling credit risk, thus becoming the main technique used for credit scorecard construction (Thomas et al., 2002).

The objective of logistic regression is to construct a function which determines the probability of default for a given client. It considers $V$ different variables (regressors) in a vector $\mathbf{x}_i \in \mathbb{R}^V, i = 1,...,N$, and an observed binary variable $y_i = 1$ if borrower $i$ defaults, and $y_i = 0$ else. Considering the dependent variable as latent, the probability of default is:

$$p(y_i = 1|\mathbf{x}_i) = \frac{1}{1 + e^{-\left(\beta_0 + \sum_{j=1}^{V} \beta_j x_{ij}\right)}}, \tag{1.1}$$

where $\beta_0$ is the intercept, and $\beta_j$ is the regression coefficient associated with variable $j$. Since these parameters are unknown, they have to be estimated using, e.g., a maximum likelihood algorithm, which results in unbiased, asymptotically normally distributed estimators $\hat{\beta}_j$ (Hosmer and Lemeshow, 2000). The expression in the exponent is a measure of the total contribution of the independent variables used in the model, and is known as the logit (Greene, 1993).

### 1.1.6 Model Validation and Implementation

The final step for every project is implementing the results of the model in a corporate environment, formalizing obtained knowledge, summarizing all information, and applying the desired results. In this phase the implementation must be planned by describing what and who shall participate in the deployment process, describing whether position responsibilities must be redefined, identifying the necessary training required for the correct performance according to these responsibilities, and discussing the methodology to keep the models up-to-date and correctly integrated with the corporate environment.

Another step is to design an evaluation methodology of the described models, measuring performance and actions against the inevitable loss of predictive capabilities that will occur if the model is not updated. In credit scoring, some measures that can be used to control the performance of the model include reports of mobility among states of arrears divided by score (vintage reports), reports on discriminating capacity of the models in each month, performance of the model for each monthly panel of loans, and measurements of the K-S statistics for the whole portfolio each period.

In general, a large quantity of new, detailed knowledge is obtained when developing a statistical model. This knowledge alters the functioning of the model itself, since at each step of the development process new information is added that can be used to improve further results, and after deployment, this do not change, with the knowledge obtained when applying the model resulting in new operational processes that enrich future developments further down the road.

### 1.1.7 Recent Developments in Credit Scoring

As was mentioned above, research in credit scoring is ample, covering widespread topics. A short survey of recent developments is reviewed in this section, showing some of the challenges that this thesis also tackles. There are three wide branches of development in credit scoring: the improvement of models by using new methodologies, applying credit scoring in different areas, and improving existing models with new techniques.

Regarding improvement of existing models, one example is the problem of imbalanced datasets that arise when there are few defaulters (non-payers), such as in loans to large companies, banks, sovereign instruments, and some categories of specialized lending, according to Van Der Burgt (2008). Recently, Brown and Mues (2012) tested multiple datasets and models and concluded that there is a need for advanced ensemble models if best results are to be obtained. This work uses a different branch of data mining, semi-supervised models, to extend existing results, as Chapter 2 and Chapter 3 show.

New areas of research in credit scoring include credit scoring for micro, small, and medium companies. This segment requires the use of statistical-based credit scoring, since its number is constantly growing, but its differences from natural persons in terms of size, income, and organizational structure lead to different characteristics and thus different types of risk measurement. Some recent examples of these efforts are, for example, the publications by Kim and Sohn (2010), or Van Gool et al. (2011), focusing mainly on small companies from the technology sector and startups. Appendix A shows our own results in credit scoring for micro entrepreneurs.

Improving results using new methodologies is always an important development, and this thesis focuses on this area exactly. One active issue is survival analysis, that is, extending the logistic methodology presented in section 1.1.5.1 to models that also include the time until defaults occur. One example of this is the work of Bellotti and Crook (2008), who studied default in credit card debt using macro-economic variables and survival analysis, and Tong et al. (2012) who studied the use of a combination of models known as the mixture cure model. This thesis expands this literature in Chapter 4. Finally, other experimental models for credit scoring were studied in Setiono et al. (2009), who used adapted neural networks – to satisfy legal constraints of credit scoring models – and applied them in consumer lending datasets.

As shown, credit scoring tries to answer many different questions, such as: Will someone default? Which variables predict default? And when will someone default? This work is driven by two more questions: Why does default occur? And, can we use the answers to these questions to improve credit scoring? The next sections explain the objectives and the publications that tackle these questions.

## 1.2 Objectives

### 1.2.1 General Objective

The main goal of this thesis is to contribute to the measurement of consumer credit risk by extending common, data mining based, credit scoring models by including the explicit strategic relationships that occur between borrowers and lenders, using game theory. This will be done applying state-of-the-art semi-supervised data mining techniques to the classical credit risk models, integrating a strategic view of the consumer-lender relationship with operational processes.

### 1.2.2 Specific Objectives

The specific objectives of this thesis are:

1. To create a theoretical framework that characterizes the current state-of-the-art in data mining techniques used for consumer credit scoring, and of the game theory techniques that can be applied to study this phenomenon.
2. To develop novel models that use an integrated vision of data mining and game theory which present significant improvements over current credit scoring systems.
3. To benchmark obtained models in real-world databases to evaluate the feasibility of the proposed approach, and its benefits/drawbacks.

## 1.3 Structure

This thesis consists of two publications plus a working paper presented at the Credit Risk Centre conference last year, and which is ready to be submitted, in the area of consumer credit scoring. Broadly, the first paper corresponds to a semi-supervised model that allows dividing defaulters into two classes: those who default due to lack of capacity to repay, and those who default due to lack of willingness to repay. The definition of defaulting due to capacity and willingness is common when describing classification variables, but this is the first time a formal model actually includes this information for defining the objective variable. The second paper goes into technical detail regarding the semi-supervised clustering technique that was used to solve the problem of the first publication, since available clustering techniques were not appropriate for the combinatorial problem that arises. The third and final publication seeks to apply the developed technique for survival analysis, studying the properties necessary for multi-class classification, and applying the technique of the first two works into a new dataset.

### 1.3.1 Improving Credit Scoring by Differentiating Defaulter Behaviour

This paper, accepted for publication at the Journal of the Operational Research Society, corresponds to the first published attempt at formalizing a well-known occurrence in credit scoring: that there are different classes of defaulters. In particular, it is known that some variables used to predict default do so because they foresee tight cash flows, resulting in lack of capacity to repay obligations. Examples of these variables correspond to the leverage of an individual (total debt divided by total income), or monthly burden (total instalments paid as a percentage of income). Other variables, in contrast, model the willingness to repay obligations, resulting in lack of willingness being defined not as a malicious act, but as behavioural patterns that lead to repayment failure. Examples of these variables include age (young individuals are more prone to take on responsibilities they cannot afford later on), or the so-called 1-2-3 arrears, the number of previous instalments paid one, two, or three days after the due date of the obligation, which indicates weakness in financial responsibility.

The formalization of this concept is performed with the aid of game theory. A very simple, but descriptive, game is modelled, in which consumers are differentiated in two ways: A borrower who fails due to lack of capacity to repay has a long planning time (modelled using different discount

factors), and desires future loans, whereas a borrower who fails due to lack of willingness to repay has a very short planning time, and is indifferent to the opportunity of new loans. The lender, on the other hand, does not know the class of each potential borrower, but uses collected information (operational databases, public databases, and the borrower's request form) to decide whether to grant or reject a given request. When all these theoretical behaviours are combined, a rational zone of granted loans is obtained, depending on each amount requested and all the conditions that were observed. This zone is described by a series of constraints for borrowers, that all of them must satisfy, thus arriving at a combinatorial clustering problem. The CCF method (explained in Chapter 3) enables solving this problem, obtaining two different classes of defaulters.

With two classes of defaulters, the credit scoring problem becomes a multi-class one, because potential clients must be classified into either good borrowers, or defaulters of one of the two classes. The last part of the paper describes the results of contrasting a binary classification approach with the multinomial one, observing an improvement of five to seven percent in classification capacity, and a much more fine-grained description of the reasons that lead to default. This improvement is significant monetarily, since better classification is achieved, and also operationally, since the new knowledge can be used for designing better origination policies, specifying what information to request when clients are applying for loans.

### 1.3.2 Semi-Supervised Constrained Clustering with Cluster Outlier Filtering

In order to solve the problem presented in Chapter 2, it is necessary to group defaulters using a series of constraints that depend on the elements that are present in each group. Moving one element from one group to the other might potentially alter the conditions for all the elements in the group, thus changing the conditions for all points. Considering that typical credit scoring databases can have from several thousand to several million different cases, the algorithm that clusters these cases must run fast, and must be able to handle a large number of different constraints.

The Constrained Clustering with Filtering algorithm, published as a book chapter appearing in Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, from the Springer series Lecture Notes in Computer Science, solves this problem by adapting K-Means

clustering, a well-known Expectation-Maximization algorithm, for constrained clustering. Since the constraints can be considered external knowledge, the proposed algorithm falls into the category of Semi-Supervised methods.

The algorithm was tested both in the credit scoring dataset and also in a synthetic dataset, with excellent results. The proposed algorithm is very fast, and shows good convergence properties.

### 1.3.3 Survival Analysis for Credit Scoring with Multiple Types of Defaulters

The final chapter looks at answering the next question in credit scoring problems: when will borrowers default? As was mentioned, this question requires the use of survival analysis techniques, a model imported from medical analysis with linear properties that make it applicable for credit scoring. This paper studies how to extend survival analysis when in the presence of more than two classes of defaulters.

In particular, two approaches are studied in the work: the first is the common competing risk applications, in which each class of event (i.e. default) is considered fully independent of the others, which translates in the credit scoring case to the fact that borrowers can simultaneously be at risk of failing because of lack of capacity to repay and lack of willingness to repay. This approach is very simple to solve, but its assumptions are highly questionable.

The second approach relaxes the independence conditions, assuming that the class of the borrower is unique, but unknown. When this assumption is made, mixture models are better suited for modelling the new situation. A mixture model, by definition an ensemble of functional forms, assumes a distribution for each class, built from the database, and different survival curves for each. In this case, the chosen models were survival analysis in Cox regression form and logistic regression, since their properties fit the legal requirements of credit scoring systems.

The methods are then benchmarked in a new dataset of loans, applying the CCF algorithm and performing benchmark classification. Again the results show that the multi-class model is dominant over binary results, and better characterizes the event of default as well. Finally, in order to allow replicability of the proposed models, an open source package is made publicly available with all the code necessary for running the CCF algorithm and efficiently solving the problem of mixture,

multi-class, survival analysis.

## 1.4 Full Contributions List

The following is the full list of indexed academic contributions made during my studies for the PhD, from March, 2009 to September, 2012. It includes indexed and peer-reviewed journal papers, book chapters, and conference papers, as well as conference presentations. Only accepted, published, or personal presentations are included.

### 1.4.1 Indexed Journal Papers

- Bravo C., Thomas, L. C. and Weber, R (2012). "Improving Credit Scoring by Differentiating Defaulter Behaviour". Journal of the Operational Research Society. Accepted for Publication. Indexed in Science Citation Index. Part of this thesis as Chapter 2.
- Brown, D., Famili, F., Paass, G., Smith-Miles, K., Thomas, L. C., Weber, R., Baeza-Yates, R., Bravo, C., L'Huillier, G., and Maldonado, S (2011). "Future Trends in Business Analytics and Optimization". Intelligent Data Analysis 15: 6. 1001-1017. Indexed in Science Citation Index Expanded.
- Bravo, C., Lobato, J.L., L'Huillier, G., and Weber, R (2010). "Probability Estimation for Multiclass Problems Combining Support Vector Machines and Neural Networks". Neural Networks World. 20(4): 475-. Indexed in Science Citation Index Expanded.
- Bravo, C., Maldonado, S. and Weber, R (2009). "Seguimiento en Modelos de Regresión Logística" [Model Follow-Up in Logistic Regression Models. In Spanish]. Revista de Ingeniería Industrial, 8(2): 31-44. Indexed by EBSCO Host.

### 1.4.2 Book Chapters and Conference Papers

- Bravo, C. and Weber, R (2011). "Semi-Supervised Constrained Clustering with Cluster Outlier Filtering". Lectures Notes in Computer Science 7042. 347-354. Indexed by Springer-

Link. Part of this thesis as Chapter 3.

- Bravo, C., Figueroa, N. and Weber, R (2010). "Modeling Pricing Strategies Using Game Theory and Support Vector Machines". Lecture Notes in Artificial Intelligence: Advances in Data Mining. Vol. 6717: 323-337. Indexed by SpringerLink.

- Bravo, C., Figueroa, N. and Weber, R (2010). "Game Theory and Data Mining Model for Price Dynamics in Financial Institutions". Proceedings of the International Joint Conference on Neural Networks (IJCNN 2010). 1-8. Indexed by Thomson Reuters ISI Proceedings, IEEE Xplore.

### 1.4.3 Conference Presentations

- Session chair, session "Risk Analysis and Credit Scoring". 26th European Conference on Operational Research. Vilnius, Lithuania. 2012.

- Bravo, C., Thomas, L. C. and Weber, R. "Improving Credit Scoring by Differentiating Defaulter Behavior".26th European Conference on Operational Research. Vilnius, Lithuania. 2012.

- Mora, C., Bravo, C. "Ensemble Methods for Classification of Volcano - Seismic Signals". 26th European Conference on Operational Research. Vilnius, Lithuania. 2012.

- Pizarro, J., and Bravo, C. "Comparative Analysis of Dynamic Models Specification for Behavioral Scoring in a Microfinance Portfolio". 26th European Conference on Operational Research. Vilnius, Lithuania. 2012.

- Bravo, C., Thomas, L. C., and Weber, R. "Competing Risks in Credit Scoring Using Survival Analysis and Economical Modeling". CRC Credit Scoring Conference. Edinburgh, United Kingdom. 2011.

- Maldonado, S., Bravo, C., and Weber, R. "Methodologies for Cut-Off Point Determination in Credit Scoring Models for Not-for-Profit Governmental Institutions". CRC Credit Scoring Conference. Edinburgh, United Kingdom. 2011.

- Bravo, C., Figueroa, N., and Weber, R. “Game Theory - Data Mining Model for Price Dynamics in Financial Institutions”. ALIO - INFORMS Joint International Meeting. Buenos Aires, Argentina. 2010.
- Bravo, C. “The Effects of Human Intervention in Statistical Models”. ALIO - INFORMS Joint International Meeting. Buenos Aires, Argentina. 2010.
- Bravo, C., Maldonado, S., and Weber, R. “Seguimiento en Modelos de Regresión Logística” [Model Follow-Up in Logistic Regression Models]. VIII Chilean Congress on Operations Research, OPTIMA 2009. Chillán, Chile. 2009.
- Bravo, C., Maldonado, S., and Weber, R. “Model Follow-Up in Credit Scoring”. 23rd European Conference in Operational Research, EURO 2009. Bonn, Germany. 2009.
- Bravo, C., and Weber, R. “Modelo de Tarificación en Base a SVMs y Teoría de Juegos” [Pricing Model Based on SVMs and Game Theory, in Spanish]. XXII National Encounter of Educators on Operations Research, ENDIO 2009. Buenos Aires, Argentina. May, 2009.

# Chapter 2

# Improving Credit Scoring by Differentiating Defaulter Behaviour[1]

**Cristián Bravo, Lyn C. Thomas, and Richard Weber**

**Abstract**

We present a methodology for improving credit scoring models by distinguishing two forms of rational behaviour of the defaulters of loans. It is common knowledge among practitioners that there are two types of defaulters, those who do not pay because of cash flow problems ("Can't Pay"), and those that do not pay because of lack of willingness to pay ("Won't Pay"). This work proposes to differentiate them using a game theory model to describe their behaviour, and then represent this separation of behaviours by a set of constraints that form part of a semi-supervised constrained clustering algorithm, constructing a new target variable summarizing relevant future information. Within this approach the results of several supervised models are benchmarked, in which the models deliver the probability of belonging to one of these three new classes (good payers, "Can't Pays", and "Won't Pays"). The process brings significant

[1]The following is an authorized copy of the paper accepted for publication at the special issue on Credit Risk at the Journal of the Operational Research Society. Please cite this paper as follows: Bravo, C., Thomas, L. C., and Weber, R. Improving Credit Scoring by Differentiating Defaulter Behaviour. Journal of the Operational Research Society. In press. 2013.

improvement in classification accuracy, delivers strong insights about the nature of defaulters, and could lead to improved credit scoring systems.

## 2.1 Introduction

Credit scoring (Thomas et al., 2002) is one of the most widely known applications of statistical models and data mining, whose goal is to differentiate between customers who will pay back a given loan, and those who will not. Classifying borrowers into these two groups (defaulters and non-defaulters) has been the standard approach of credit scoring from its inception. Lately there has been a rise in the number of statistical models that use economic analysis and game theory to better understand the behaviour of the relevant players in many applications. It has been used to improve manufacturing strategies (Wang, 2007), in credit card fraud detection (Vatsa et al., 2005), in marketing to improve competition capabilities (Bravo et al., 2010), and in *phishing* detection for spam filtering (L'Huillier et al., 2010), to name a few. In this paper we present a procedure that extends application credit scorecards by differentiating defaulters into two groups, those who default due to problems associated with willingness of repayment, and those who fail to pay because they do not have the capacity to do so.

The competitive relationship between lenders and their borrowers has been studied by a number of authors who model who are defaulters and what are the probabilities of default. For example in the well-known work of Stiglitz and Weiss (1981), the authors imposed conditions on the rationality of the players which produce an adverse selection game in the credit granting process. The importance of collateral as a way of selecting customers was also studied by Wette (1983). Both papers point out that the drivers for the decisions of whether to request a loan and whether to grant the request include the amount of the loan and the collateral offered.

More recent works have focused more on reasons and characteristics of default. Alary and Gollier (2004), using collateral and amount lent as their main decision variables, emphasized the moral hazard problem that lenders faced, and showed that under certain conditions customers will default strategically. Moffat (2005) used hurdle models to model default, and the extent of it, also finding different levels or intensities of default. Block-Lieb and Janger (2006) suggested that the expectation that borrowers will be fully rational is not true in some cases, and that it is not

reasonable to assume that all customers behave strategically. The work of Guiso et al. (2010) also focused on the reasons that drive default, and found that there are segments of the population that default strategically, and that such behaviour is driven by different economic and moral variables. Finally, differences in default and the characteristics of defaulters have also been studied in small business lending, with Lin et al. (2011) defining four different types of default depending on the financial conditions of borrowers.

There is agreement in the literature that there are different reasons for defaulting even though credit scoring aggregates these different reasons into just one default class. In credit scoring default is normally defined as being 90 days in arrears, and one normally estimates the risk of this happening in the next year. Our proposal is to differentiate between two types of default, for different reasons, by using first economic modelling, and then semi-supervised clustering. Then it is possible to build a scoring system using supervised techniques such as logistic regression and neural networks to predict whether a borrower will default for each specific reason.

This paper is divided as follows: first we present a game theory model that determines which combination of collateral, loan rate charged, and borrower/lender characteristics lead to loans being given, but then defaulted. The borrowers are assumed to belong to two groups; those who are unwilling to repay the loan (referred to colloquially as “Won’t Pays”) and those who want to repay but may default because they do not have the capacity to repay (called colloquially the “Can’t Pays”).

The second part of the paper reports using this expected rational behaviour described to create a semi-supervised clustering methodology, that determines which borrowers are more likely to present which behaviour. In order to demonstrate the usefulness of the proposed model, the resulting procedure is applied to a dataset of actual loans in the Validation and Experimental Results section. In it, the clustering procedure is conducted to find which are the “Won’t Pay” and which the “Can’t Pay” clusters. Then, two supervised learning models (neural networks and multinomial logistic regression) are benchmarked against the classical logistic regression model to obtain an overall estimation of the credit risk of the borrowers, focusing on both the gains in accuracy and in obtained knowledge. Finally the conclusions extracted from building such a model are outlined.

## 2.2 Game Theory Model of the Loan Granting Process

Throughout the work we will assume there is set of $N$ borrowers that were granted a loan. Each loan is described by $V$ different variables, stored in database $X$, representing the different characteristics of the loan, the borrower, and the evolution of the loan. There is also an outcome variable, given by $d_i \in \{0,1\}$, associated with each $x_i \in X$ ($i \in \{1,\ldots,N\}$), indicating whether that borrower defaulted on that loan or not. Only the characteristics in $V$ which describe data known before the loan was granted can be used for classification, which we denote as $X_{past} \subset X$. The remaining characteristics, $X_{future} \subset X$, consist of information that can only be used for determining whether and for what reasons the borrower has defaulted. We require that $X_{past} \cup X_{future} = X$, and that $X_{past} \cap X_{future} = \emptyset$, in order to avoid bias in the classification procedure. Since application credit scoring is aimed at new customers, the database will consist of different applicants, with one loan per applicant.

In general, a game is defined by its $N$ players, their strategies $s_i$, and the payoffs $u_i(s_i, s_{-i})$ of playing the various strategies in the competitive or adversarial setting (Fudenberg and Tirole, 1991), where $s_{-i}$ represents the strategies played by all other players except $i$. In the model of the loan granting process presented in this work, the first player, $L$, is the lender who wants to maximize the profit of lending to $N$ potential borrowers over two time periods. Each borrower has an income of $I_i$ given at the end of each period, provided they have not previously received a shock to their cash flow (for example, become unemployed) which can happen with probability $q$ during each period.

The borrowers can be of two types, $C$ or $W$, where $C$s are the "Can't Pays" who are willing to pay back the loan they received provided they are able, and so will only default if they cannot pay because of the occurrence of the external shock. Class $W$ corresponds to the "Won't Pays" who are unwilling to pay back the loan even if they can afford to do so, and so presumably took the loan with no intention of paying it back. We can also identify the good payers (class $P$) as members of class $C$ that are not subject to an external shock, and thus were able to repay the loan.

The mechanism for granting loans will be the following: in the first period borrower $i \in \{1,\ldots,N\}$ can ask for a loan $x_{i1}$, and the lender can agree to or refuse this request, where $y_{i1} = 1$ is agreeing to the loan and $y_{i1} = 0$ is refusing it. The loan is paid at the beginning of the period, and an interest rate of $r$ is charged with the loan as the interest to be paid back at the end of the period. The lender will also ask for collateral $C_{i1}$ from the borrower $i$ to secure the loan, where $C_{i1}$

could be, for example, the deed to the house on which a mortgage is taken out; or the lender could grant the loan without collateral (an unsecured loan) when $C_{i1} = 0$. If the loan is not repaid, the lender will sell the collateral but will only get an amount $\alpha C_{i1}$ back, with $\alpha \in [0,1]$, so $(1-\alpha)$ is the so-called "haircut" on the collateral. Borrowers of type $C$ will not be able to repay if they have had a shock to their cash flow during the period, while those of type $W$ will not pay no matter what their cash flow is.

If the loan is repaid, the borrower can repeat the process in the second period asking for a loan of value $x_{i2}$ which, if it is given, will correspond to decision $y_{i2} = 1$ by the lender, as opposed to $y_{i2} = 0$ if refused. The interest rate remains $r$, but the collateral this time would be $C_{i2}$. If the loan is not repaid in the first period, the borrower will not be allowed to apply for a new loan in the second period. If the first loan was not granted, it is not reasonable that this would be different in the second period, so it will not be granted as well.. The time dependence of the utility of the players is given by discount rates defined to be $\delta_L$, $\delta_C$ and $\delta_W$ for the lender and the two types of borrowers.

If we do not restrict the total amount lender has available for lending, then this procedure can be considered as a series of two-player games between the lender and each of the individual borrowers, with a generic income of $I$ (different) for each one. In this case, without loss of notation, we will ignore the index $i$ of each borrower. The payoff of the two types of borrowers and the lender, $u_C$, $u_W$, and $u_L$, will depend on the strategies $s = (y_1, y_2, x_1^C, x_2^C, x_1^W)$ chosen, together with the belief $\theta$ of the lender that it is facing a borrower of type $C$. The lender does not know which type of borrower is facing – this is private information of the borrower – so this is a game with incomplete information.

The expected utility of the borrowers is the net present value of their income and the loans they receive, according to their type:

$$\begin{aligned} u_C(s) =& y_1(x_{C1} - q\delta_C C_1 - (1-q)\delta_C(1+r)x_{C1}) + (1-q)\delta_C I \\ &+ (1-q)\delta_C y_1 y_2(x_{C2} - q\delta_C C_2 - (1-q)\delta_C(1+r)x_{C2}) \\ &+ (1-q)^2\delta_C^2 I \end{aligned} \tag{2.1}$$

$$u_W(s) = y_1 x_{W1} - y_1\delta_W C_1 + (1-q)\delta_W I + (1-q)^2\delta_W^2 I \tag{2.2}$$

where the last two terms of $u_W(s)$ correspond to the expected income at the end of the first and second periods.

The utility of the lender is the expected value of the returns from the loan:

$$\begin{aligned} u_L(s) =& y_1(1-\theta)(-x_{W1}+\delta_L\alpha C_1) \\ &+ y_1\theta(-x_{C1}+q\delta_L\alpha C_1+(1-q)\delta_L(1+r)x_{C1}) \\ &+ \theta(1-q)\delta_L y_1 y_2(-x_{C2}+q\delta_L\alpha C_2+(1-q)\delta_L(1+r)x_{C2}) \end{aligned} \tag{2.3}$$

To derive the set of conditions that the player must fulfil, we will create a set of restrictions that account for the individual rationality of the players in general, that is, we will demand that the best possible choice is to request and grant two loans, which is the scenario that we propose represented reality best since we observe the loans in the database.

We will require one condition for this setting: that the strategy of granting or requesting two loans (or one in case of borrowers in class $W$) is individually rational, so requesting and granting two loans provides positive utility, as opposed to not participating or to just requesting a loan one period for class $C$, which would bring a utility given by the income they should receive on the periods without a loan request.

For the lender, the conditions are:

$$u_L(y_1=1,y_2=1,C_1,C_2,s'_{-i}) \geq \quad u_L(1,0,C_1,0,s'_{-i}) \quad \forall C_1,C_2 \tag{2.4}$$

$$u_L(1,1,C_1,C_2,s'_{-i}) \geq \quad u_L(0,0,0,0,s'_{-i}) \quad \forall C_1,C_2 \tag{2.5}$$

The right-hand side of expression 2.4 is equal to:

$$\begin{aligned} u_L(y_1=1,y_2=0,C_1,C_2,s'_{-i}) =& \theta x_{C1}(-1+\delta_L(1-q)(1+r)) \\ & -(1-\theta)x_{W1}+\delta_L\alpha C_1(1-\theta+\theta q) \end{aligned} \tag{2.6}$$

And so expression 2.4 is equivalent to:

$$x_{C2}(-1+\delta_L(1-q)(1+r))+(1-q)q\delta_L\alpha C_2 \geq 0 \tag{2.7}$$

Expression 2.5 is equivalent to:

$$\begin{aligned} &\theta x_{C1}(-1+\delta_L(1-q)(1+r))+x_{C2}\delta_L\theta(1-q)(-1+\delta_L(1-q)(1+r)) \\ &-(1-\theta)x_{W1}+C_1\alpha\delta_L(1-\theta(1-q))+\theta(1-q)q\delta_L^2\alpha C_2 \geq 0 \end{aligned} \tag{2.8}$$

To obtain the solution space, the same analysis has to be carried out for each of the consumers. For consumers of class $C$, the equations that must be fulfilled are:

$$u_C(x_{C1}>0,x_{C2}>0,s'_{-i}) \geq u_C(x_{C1}>0,x_{C2}=0,s'_{-i}) \tag{2.9}$$

$$u_C(x_{C1}>0,x_{C2}>0,s'_{-i}) \geq u_C(x_{C1}=0,x_{C2}=0,s'_{-i}) \tag{2.10}$$

Note that this assumes that $y_1=1$ and $y_2=1$, i.e. equations (2.4) and (2.5) are satisfied. This is reasonable because every other decision implies utilities of $u_C(x_1=0,x_2=0,s'_{-i})$, that is, no loans are given.

The RHS value, from (2.10) is:

$$u_C(x_1=0,x_2=0,s'_{-i}) = I(1+\delta_C(1-q)+\delta_C^2(1-q)^2) \tag{2.11}$$

So, the borrower only asks for a loan if the utility is greater than the expected value of the

income he or she will receive. Incorporating (2.11) in (2.10) implies:

$$\begin{aligned} x_{C1}(1-\delta_C(1-q)(1+r))+\delta_C x_{C2}(1-q)(1-\delta_C(1-q)(1+r)) \\ -\delta_C q C_1-\delta_C^2(1-q)qC_2 \geq 0 \end{aligned} \tag{2.12}$$

Condition (2.9) assumes that the expected utility that arises from applying for a second loan must be positive, so that:

$$x_{C2} \geq \frac{\delta_C q}{1-\delta_C(1-q)(1+r)} C_2 \tag{2.13}$$

Finally, borrowers of class $W$ will apply for a loan if their utility is greater than the discounted collateral they would lose if they applied for the loan:

$$x_{W1} \geq \delta_W C_1 \tag{2.14}$$

The set of inequalities (2.7), (2.8), (2.12), (2.13), and (2.14), creates a space of behaviours that is individually rational. This set of equations can be interpreted as a set of constraints that will be used to separate the behaviour of the defaulters. A defaulter of class $C$ desires a second loan even though this will be refused since he or she will have defaulted, and a defaulter of class $W$ does not expect a second loan. Given this set of constraints, it is now possible to actually differentiate the defaulters: we can obtain two groups that are similar (in variance) and fulfil the restrictions that are proposed here.

Some other interesting results that arise from the proposed model are:

- Since the LHS of (2.5) is increasing in $C_1$, $C_2$, and $\delta_L$ the collaterals are an incentive for the lender to grant loans. Similarly, a lower expected return on capital, and so a higher discount factor $\delta_L$, increases the chance the loan will be given.
- In (2.13) and (2.14) the model reflects that borrowers might accept a loan even if the collateral

demanded is of greater value than the loan itself. Such event is observed in real life, for example, when a loan to finance a percentage of a car is secured by the value of the car in full.

- A necessary condition for the set defined by the constraints to be non-empty is that $\delta_W < \delta_C < \delta_L$. This is reasonable since people who need to borrow have lower discount factors than the lending organizations, and borrowers in class $W$ are assumed to be more short-sighted than regular borrowers. A high value of $\delta_W$ would also result in that group being limited to very high loan amounts (relative to the collateral), or to borrowers with no collateral. Allowing borrowers in class $W$ to be more short-sighted than borrowers in class $C$ gives a wider range of options for both groups.

Now it is possible to propose a learning model that incorporates such restrictions, as will follow in the next section.

## 2.3 Constrained Clustering and Semi-Supervised Methods

Constrained clustering (Basu et al., 2008) is a semi-supervised approach to obtaining segments of a dataset incorporating certain restrictions that must be fulfilled by the members of the cluster, the members of different clusters, or the general structure of the clusters. The term "semi-supervised" refers to the incorporation of knowledge that is not directly present in the data – or that is known only for a limited number of cases – in order to improve the results on the whole domain. In this particular case, restrictions are added to a clustering procedure so that the objective is not only to minimize intra-cluster variance, but also to satisfy a set of conditions for each member, or each cluster. The methodology has been applied successfully in several fields, ranging from signal processing (Levy and Sandler, 2008), epidemiology (Patil et al., 2006), to OR applications (Bard and Jarrah, 2009).

There are two different approaches for constrained clustering, as noted by Davidson and Ravi (2005), and both are based on the concept of "Must-Link" and "Cannot-Link" constraints. The

first set of restrictions indicate that two elements must be in the same cluster, whereas the second set prohibits the presence of two elements in the same cluster. The methods differ in the role of restrictions: in the first case the algorithm fulfils an objective or distance functions using the information from the constraints. The best known application of this work is by Basu et al. (2004). In the second case the constraints simply limit the presence of elements in the same cluster. Our work is framed in the second type of applications, with a minor adjustment that adds much complexity: the elements in one cluster must satisfy the restrictions against most of the elements of the other cluster. There are algorithms that solve this particular problem, such as the CCF algorithm presented by Bravo and Weber (2011), which will be used in this paper.

### 2.3.1 A Model of Constrained Clustering to Separate Classes of Defaulters

The constraints obtained from the economic assumptions from previous sections can now be used to design a new objective variable using semi-supervised methods. The overview of the process is as follows:

1. Select defaulters from the database $X$, and describe them using only the information from variables in $X_{future}$, that is, variables collected after the loan has been granted. Call this $X_d$.
2. Cluster elements into two groups, one for each type of defaulter ($C$ and $W$) using the CCF algorithm with dataset $X_d$ as input. The constraints defined in the previous section are extended now to the whole cluster, requiring that cases in cluster $W$ have to satisfy their own condition, equation (2.14), elements in cluster $C$ must satisfy constraint (2.7) and constraint (2.10), and there is a cross-cluster constraint, given by the lender condition (2.8) which must be satisfied by most pairs of elements in different clusters.
3. With the elements clustered, the new objective variable is simply extending the default variable to the new case when there are three different cases: good payers, defaulters in class $C$, and defaulters in class $W$.

The exact constrained clustering problem to be solved considers two groups with centroids given by $c_k$, $k = \{1,2\}$ and a binary variable $m_i$ for each $i \in \{1,\ldots,N\}$ that represents the class

of the borrower (1 for class $C$, 0 for class $W$). Since the value of the second loan and the second collateral are not known, we will assume they are a fraction of the original value requested ($f_{Cr}$ and $f_{Col}$), as explained below. Since now each borrower has his or her own variables, we will refer to the amount borrowed as $x_i^a$, to the collateral given for the first loan as $C_i$ and finally to the interest rate paid as $r_i$.

The problem is then to solve the following optimization problem, adapted from the formulation presented by Doğan and Güzeliş (2006):

$$\min_{m_1,m_2,c_1,c_2,max_W,min_C} \sum_{i=1}^{N} m_i\|x_i-c_1\|^2+(1-m_i)\|x_i-c_2\|^2 \tag{2.15}$$

$$\begin{array}{llll}
s.t. & min_C & \geq & (1-\theta)x_i^a \quad \forall i \mid m_i=0 \\
& x_i & \geq & \delta_W C_i \quad \forall i \mid m_i=0 \\
& 0 & \leq & x_i^a(1-\delta_C(1-q)(1+r_i))(1+\delta_C f_{Cr}(1-q)) \\
& & & -C_i\delta_C q(1+\delta_C f_{Co}) \quad \forall i \mid m_i=1 \\
& 0 & \geq & \theta x_i^a(-1+\delta_L(1-q)(1+r_i))(1+f_{Cr}\delta_L(1-q)) \\
& & & +C_i\alpha\delta_L(1-\theta(1-q)q\delta_L f_{Co})-(1-\theta)max_W \quad \forall i \mid m_i=1 \\
& min_C & \leq & \theta x_i^a(-1+\delta_L(1-q)(1+r))(1+\delta_L f_{Cr}(1-q)) \\
& & & +C_i\alpha\delta_L(1-\theta(1-q)+\theta f_{Col}(1-q)q\delta_L) \quad \forall i \mid m_i=1 \\
& max_W & \geq & x_i^a \quad \forall i \mid m_i=0 \\
& c_1 & = & \frac{\sum_{i=1}^N m_i x_i}{\sum_i m_i} \\
& c_2 & = & \frac{\sum_{i=1}^N (1-m_i)x_i}{N-\sum_i m_i} \\
& m_i & \in & \{0,1\}
\end{array} \tag{2.16}$$

From (2.15) it can be derived that this is a quadratic problem with integer variables, which is difficult to solve, making a heuristic approach necessary. Furthermore, the constraints (2.16) depend directly on the cluster in which the elements are present, so an extensive expression (one that does not include the $m_i$ parameters in the description) would have to include $N \times N$ restrictions, one for each pair of elements, turning the problem intractable for large databases.

The CCF algorithm follows a similar methodology as the K-Means algorithm, extending the procedure for constrained problems. In each iteration all constraints are checked for each element, i.e., if the element is in class $C$, then equations (2.12) and (2.13) are evaluated for the element itself,

and equation (2.8) is evaluated against the extreme elements of cluster $W$. In case the conditions are not fulfilled, the element is changed from cluster, and if that does not solves the issue, the extreme values in both clusters are removed from the analysis and the process is repeated. The algorithm continues until the violations are below a threshold and the centroid values do not move more than a given tolerance.

The elimination of extreme values in each iteration gives a relaxation of the problem. The elements in each cluster have to satisfy the constraints against most of the elements in the other cluster, and this is accomplished by eliminating a small number of extreme cases in each iteration and ensuring that all the rest of the cases satisfy the constraints. In the end, since the final cluster of the element is calculated only by proximity to centroids $c_1$ and $c_2$, it is possible to assign even the cases that were eliminated from the clustering procedure.

One of the open questions that remain is what are the possible values of the parameters used in the modelling. We propose the following values, which correspond to realistic measures in credit risk:

- $q$: This parameter represents the chance that a borrower receives a shock to his or her income, and so is forced to default. This value could correspond to the long term default rate for loans, because it is expected that in the long term, most of the "Won't Pay" borrowers are filtered from the database. Another option (and the one used in the experimental results section) is to discount the observed default rate ($DR$) by a small amount, in order to include the unobserved segment. A possible value would be $q = DR \cdot \frac{1-\theta}{2}$.
- $\theta$: This parameter represents the *a priori* probability of a customer being in the class "Can't Pay". The lender can estimate this value from historical data, since the number of customers who, after default, paid back very little of the loan or nothing at all can be used as an estimation of the number of customers in class $W$. In the experimental part of this paper the proportion of defaulters among all defaulters that made a payment up until two years after the default occurred was used as an approximation of this parameter.
- $\alpha$: The expected recovery to be extracted from the collaterals is a known value to credit granters, with values commonly between 40% to 60% of the collateral value (see for example Yamashita and Yoshiba (2010) or Jokivuolle and Peura (2000)).

- $\delta_L$: The value of the discount factor for a company should be a known value, e.g. the expected internal return rate for the company fiscal year.

- $\delta_C$ and $\delta_W$: The value of the discount rate for the customers is more complicated to determine. Discount factors in persons have been studied on several occasions with widely different results. For example Burks et al. (2008), Chabris et al. (2008), and Green et al. (1994) report discount rates ranging from 0.1 to 0.9 depending on several factors, although these studies do not focus on financial decisions. The well-known paper of Benzion et al. (1989) gives discount factors depending on the financial amount at risk that ranges from 0.2 to 0.75, with a strong dependence on both the amount at risk and the duration of the loan. We propose an exogenous measure, for example, the maximum interest rate allowed for loans, which represents the maximum discount rate that any customer is legally bound to pay, fixing that value for $\delta_C$. As for $\delta_W$, the value has to be lower than that associated with customers of class *C*, because their demand for the loan is more immediate. We present a range of values in $[0, \delta_C]$ in the Experimental Results section.

- Values for the second loan: Two of the parameters that must be decided are the values for the second loan and the second collateral. Since defaulters were not allowed to take a new loan, these values are unknown, but they are known for the customers who successfully paid back their loans. Good proxies for these values are the proportions between the amounts that were requested/granted by returning borrowers, that is, the average value of $f_{Cr} = \frac{\texttt{Second\_Amount}}{\texttt{First\_Amount}}$ and $f_{Col} = \frac{\texttt{Second\_Collateral}}{\texttt{First\_Collateral}}$ calculated for borrowers who successfully returned the loans that were granted to them.

## 2.4 Validation and Experimental Results

In this section we present the results of applying the proposed approach on a dataset of loans granted at a local financial institution. First the dataset is introduced, then we present the procedure of constrained clustering on the defaulters from this dataset. In section Clustering Procedure we perform sensitivity analysis and selection of the parameters in the models. Finally, we analyse in detail the obtained results.

### 2.4.1 Available Dataset

A dataset consisting of 97,254 loans granted to mass-market, low and middle income independent borrowers (USD 300/month to USD 3,000/month) was used. The database originates from a Chilean organization, and comprises an 11-year period, from 1997 to 2007. The dataset has a default rate of 25.2%, with each loan described by 25 different variables representing both the customer (socio-demographic variables) and the loan itself. As was mentioned earlier, the variables describe either the granting process, that is, information that may be used for classification, or the evolution of the loan, information which can only be used for designing the target (objective) variable, which we use for the clustering procedure previously introduced. The variables available for classification, i.e. the ones obtained before granting the loan, correspond to:

- Economic Activity: The sector of the economy in which the customer is involved (through his/her job or company). The large number of sectors was clustered to improve interpretation of the variable, mapping the 47 different sectors to three larger groups. The variable was then transformed into dummies (Activity_A, Activity_B, and Activity_C), from which the last one was selected as a reference category in order to avoid multicollinearity in the models.
- Ownership of Housing: This shows if the customers own, rent or hold other types of agreements on their current home. Four classes are recognized: Owner, Tenant, Share-Tenant (Share), or other types. The class “Others” is used as a reference.
- Number of Properties: The number of properties the borrower possesses. The variable was divided into three categories: No properties, one property, or two or more properties, which was used as a reference category.
- Region of country: Division of the country into three regions.
- With Guarantor: Whether the customer has a guarantor for the loan or not.
- Length of loan: The length of the loan requested by the borrower. This number is determined by the customer, with the company simply granting the loan or refusing it, so it is not susceptible to manipulation. The durations of the loans are between one and twelve months.
- Age: The age of the customer in years.

The previously presented variables describe the customer at the moment of requesting the loan. On the other side, the variables that describe the actual evolution of the loan are:

- Collaterals: The collaterals are described by two variables. The first is a dummy variable that describes whether the customer had to give collateral on the loan (With_Collaterals), and the value of the collateral (Value_Collateral_UF). The latter is given in "development units" (*Unidades de Fomento*, referred to as UF for their acronym in Spanish), the Chilean inflation-indexed unit, that is equivalent to roughly 46 USD.
- Amount and Rate: The amount of the loan, in UF, and the total annual interest rate charged for the loan.
- Days in arrears before defaulting (Days_Arrear): The total sum of days the instalments of the loan were in arrears before defaulting. For example, if a loan defaulted in the fourth instalment, and the first was paid 10 days late, the second on time, and the third 45 days late, the variable has a value of 55.
- Cancellations: Sometimes the institution will cancel the payment of penalties and excess interest that arises from arrears, upon agreement that the next instalment is paid on time, or that a renegotiation is performed. This event is summarized into two different variables, considering the number of times this happened in the lifetime of the loan (Num_Cond), and the amount that was reduced (Amount_Cond). Additionally if some of the interest due to be paid is also discounted from the instalments, this value is reported in the variable Interest_Low.
- Extensions: Sometimes the company will extend the period of an instalment for 30 days or a similar span of time, subject to adjustment in the amount owed. The number of times a customer applies for this appears in variable Num_Post, and the amount adjusted appears in Amount_Adjust, and, since the adjustment can be positive or negative, the total amount of negative adjustments is incorporated into Negative_Adj.

The mean values of all the variables, the standard deviation, and their minimum and maximum values can be observed in Table 2.1.

Table 2.1: Descriptive Statistics

| Variable | Mean | Std. Dev. | Min. | Max. |
|---|---|---|---|---|
| With_Collaterals | 0.35 | 0.48 | 0 | 1 |
| Amount | 22.21 | 19.2 | 1.4 | 111.7 |
| Days_Arrear | 75.39 | 99.57 | 0 | 508 |
| Num_Cond | 0.34 | 0.52 | 0 | 4 |
| Num_Post | 0.54 | 0.82 | 0 | 11 |
| Num_Reneg | 0.5 | 0.77 | 0 | 6 |
| Amount_Adjust | -0.09 | 2.25 | -113.3 | 143.3 |
| Amount_Cond | 2.78 | 6.59 | 0 | 126.4 |
| Negative_Adj | 0.51 | 1.78 | 0 | 113.3 |
| Interest_Low | 0.54 | 3.91 | 0 | 142 |
| Payments | 2.49 | 2.18 | 0 | 40 |
| Value_Coll_UF | 20.22 | 92.56 | 0 | 4129.9 |
| Rate | 1.1 | 0.02 | 1.01 | 1.19 |
| Activity_A | 0.22 | 0.41 | 0 | 1 |
| Activity_B | 0.35 | 0.48 | 0 | 1 |
| Owner | 0.62 | 0.48 | 0 | 1 |
| Tenant | 0.18 | 0.39 | 0 | 1 |
| Share | 0.03 | 0.18 | 0 | 1 |
| No_Property | 0.59 | 0.49 | 0 | 1 |
| One_Property | 0.26 | 0.44 | 0 | 1 |
| Region_1 | 0.11 | 0.32 | 0 | 1 |
| Region_2 | 0.28 | 0.45 | 0 | 1 |
| Guarantor | 0.11 | 0.32 | 0 | 1 |
| Term | 10 | 1.64 | 1 | 12 |
| Age | 47.61 | 13.87 | 18 | 80 |

### 2.4.2 Clustering Procedure

For the clustering procedure, the values proposed before are used as shown, together with their origins, in table Table 2.2.

The process is run using the 24,576 customers flagged as defaulters, normalizing the dataset with all the future variables. The results consist of two clusters with 4,762 customers selected in the class “Can’t Pay” and 19,814 customers in the class “Won’t Pay”. The centroids for each class are shown in Table 2.3. The variables associated with collaterals have a greater impact in the “Can’t Pay” class, even though the difference in the percentage of customers with collaterals is not huge (40% versus 33%). The value of the collaterals for class $C$ is almost nine times greater than that in class $W$, which is a direct effect of the construction procedure of the clusters since they are active

Table 2.2: Parameters selected for clustering experiment.

| **Parameters** | **Value** | **Origin** |
|---|---|---|
| q | 0.130 | Adjusted default rate. |
| $\theta$ | 0.550 | Estimated from historical return data. |
| $\alpha$ | 0.600 | Range $[0.4, 0.6]$ |
| $\delta_L$ | 0.935 | Company Factor - 10% Inflation |
| $\delta_C$ | 0.625 | 55% maximum annual rate (Central Bank) |
| $\delta_W$ | 0.500 | From sensitivity analysis |
| $f_{Cr}$ | 1.320 | Examples from good payers. |
| $f_{Col}$ | 2.470 | Examples from good payers. |

in the restrictions, but combined with the results that the percentage of borrowers with collaterals are relatively balanced between the two clusters hints that it is not whether collateral is requested for a loan what differentiate defaulters, but the value of the collateral, which is an interesting result.

The analysis of the rest of the variables bring interesting results as well. Customers in class $C$ request a far larger amount for their loans, which would be consistent with a default based on the capacity to pay. Considering the total number of days in arrears, class $C$ accumulates almost 100 days more than class $W$ before defaulting, indicating that they make a greater effort to pay back the loans than customers in class $W$ do. Also, they apply for a larger number of renegotiations (0.71 per customer on average), get greater adjustments and debt relieves, and are more susceptible to receive a discount on their interests due (1.18 UF per customer on average, versus 0.27).

The values of the variables suggest that there is, indeed, different behaviours detected. With these results we now perform a sensitivity analysis on the parameters, to then show that these behaviour patterns help to obtain better discrimination between payers and defaulters. The output from the clustering procedure is now used to construct a new target variable with the three detected classes (payers $P$, customers with problems with capacity of repayment $C$, and customers with low willingness to pay $W$).

Table 2.3: De-normalized results of semi-supervised clustering.

| Variable | Class $C$ | Class $W$ |
|---|---|---|
| With_Collaterals | 0.40 | 0.33 |
| Amount | 48.67 | 12.89 |
| Days_Arrear | 882.53 | 708.61 |
| Num_Cond | 0.27 | 0.37 |
| Num_Post | 0.84 | 0.43 |
| Num_Reneg | 0.71 | 0.43 |
| Amount_Adjunt | -0.26 | -0.07 |
| Amount_Cond | 4.02 | 2.35 |
| Negative_Adj | 0.38 | 0.08 |
| Interest_Low | 1.18 | 0.27 |
| Payments | 2.86 | 2.37 |
| Value_Coll_UF | 58.93 | 6.59 |
| Rate | 1.10 | 1.10 |

### 2.4.3 Sensitivity Analysis and Parameter Selection

In order to check the effects of the parameters of the clustering method, the procedure was run using several values of the parameters $(\delta_W, \alpha)$. The value of $\delta_W$ was tested in the range $[0, 0.6]$, following the restrictions presented in the previous section, and $\alpha$ was simultaneously tested in the range $[0.4, 0.6]$. Parameter $\alpha$ was varied in steps of 0.05, while parameter $\delta_W$ was varied in steps of 0.1.

Table 2.4: Percentage of cases in class $W$ depending on parameters $\alpha$ and $\delta_W$.

| $\delta_W$ / $\alpha$ | 0.40 | 0.45 | 0.50 | 0.55 | 0.60 |
|---|---|---|---|---|---|
| 0.0 | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% |
| 0.1 | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% |
| 0.2 | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% |
| 0.3 | 86.17% | 85.08% | 80.14% | 86.17% | 81.13% |
| 0.4 | 74.52% | 74.09% | 74.32% | 75.09% | 75.06% |
| 0.5 | 79.82% | 74.56% | 72.14% | 71.92% | 70.79% |
| 0.6 | 77.99% | 78.78% | 78.14% | 69.51% | 69.58% |

Table 2.4 presents the obtained results, and shows that when the discount factor for the "Won't Pays" is low, then the model assigns a large number of cases to that class. This is reasonable as a low value in the discount factor implies that the restriction is more relaxed, so it is easier to satisfy the restrictions for that class. Of greater interest is that this dependency presents very little variation when varying parameter $\alpha$ for all values except $\delta_W = 0.6$, which seems to imply that after a certain threshold, the restrictions tend to balance and more cases can be assigned to class $C$. Higher values of parameter $\alpha$ relaxes the constraints applied to class $C$, but this effect was not strong until the discount factor $\delta_W$ reached a value of 0.5.

Considering the results obtained, the values $\delta_W = 0.5$ and $\alpha = 0.6$ are selected. This is done because when $\delta_W = 0.5$, the numbers of elements in cluster $W$ start to decrease, and this effect stops in $\delta_W = 0.6$, when the number does not vary notably. This indicates that the value of $\delta_W = 0.5$ is a critical value for the discount factor. When $\delta_W$ is fixed, a value of $\alpha = 0.6$ presents approximately 6,600 cases in class $C$, which is a sufficient number to ensure a valid statistical model when attempting to use these results to classify them applying supervised models.

### 2.4.4 Classification Results

In order to show the potential of our proposed approach we apply two different procedures to the same dataset. First we apply standard logistic regression without differentiating defaulters, and then we apply two methods for classification with three classes ($C$, $W$, and $P$), namely multinomial regression and a feed-forward neural network with a Multi-layer Perceptron (MLP) architecture.

Multinomial logistic regression (Hosmer and Lemeshow, 2000), is the natural extension of the binomial logistic regression model, the most widely used model in credit scoring applications worldwide with an estimated 90% of credit scoring systems using this methodology (Thomas, 2000). The method estimates only two regressions, considering that the last class is a neutral one (all parameters $\beta_W = 0$), and measures the differences between the elements. That way, the results correspond to a matrix of $(V_{past} + 1) \times \{P, C\}$ that gives the coefficients associated with the two classes, assuming the neutral value of 0 for class $W$. In this case the classification function for each class $k, k = \{P, C\}$ corresponds to:

$$p_k(x) = \frac{\exp(\beta_0^k + \sum_{j=1}^{V_{past}} \beta_j^k x_j)}{1 + \sum_{k' \in \{P,C\}} \exp(\beta_0^{k'} + \sum_{j=1}^{V_{past}} \beta_j^{k'} x_j)} \tag{2.17}$$

and for class $W$ the classification function corresponds to:

$$p_k(x) = \frac{1}{1 + \sum_{k' \in \{P,C\}} \exp(\beta_0^{k'} + \sum_{j=1}^{V_{past}} \beta_j^{k'} x_j)} \tag{2.18}$$

The second multinomial model used is feed-forward neural networks, judged to be the most accurate procedure for credit scoring according to Baesens et al. (2003). Neural networks are known to be black-boxes, but several adjustments can be made to the design in order to obtain parameters that are consistent with probabilities and that fulfil the legal requirements (De Waal et al., 2005). In particular, a configuration based on linear transfer functions and *softmax* (logistic) output functions is used. In this case, a matrix of size $(V_{past}+1) \times \{P,C,W\}$ is obtained, explicitly modelling the differences between all the classes. The classification function for each $k, k = \{P,C,W\}$ corresponds to:

$$p_k(x) = \frac{\exp(\beta_0^k + \sum_{j=1}^{V_{past}} \beta_j^k x_j)}{\sum_{l \in \{P,C,W\}} \exp(\beta_0^l + \sum_{j=1}^{V_{past}} \beta_j^l x_j)} \tag{2.19}$$

The results from both models are benchmarked against the results from a regular logistic regression, which delivers a unique set of parameters $\beta = (\beta_0, \ldots, \beta_{V_{past}})$ that construct probability $p(x)$ of being a defaulter in class $D = \{C,W\}$, a unique class. The expression of $p(x)$ is:

$$p(x) = \frac{1}{1 + \exp(-\beta_0 - \sum_{j=1}^{V_{past}} \beta_j x_j)} \tag{2.20}$$

In order to construct the experiments, a balanced sample was taken from database $X_{\text{past}}$, selecting 14,286 cases representing the three classes. Then the database was split into ten pieces, so at each turn nine segments are used to train the model, and the remaining piece is used to test the

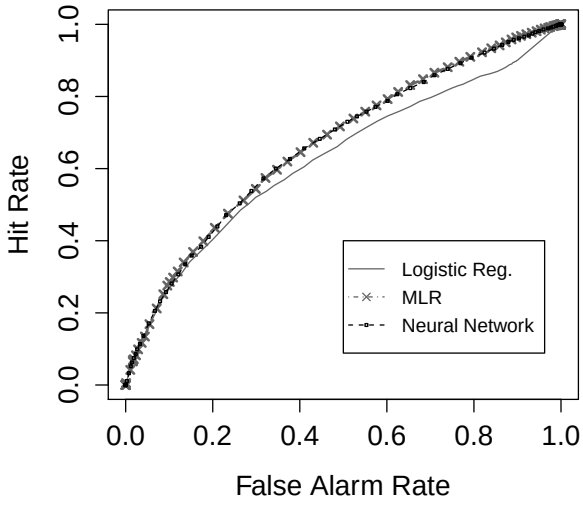



Figure 2.1: ROC curves for the three proposed models.

model. This process is repeated ten times, following a 10-by-10 cross-validation procedure (Hastie et al., 2009). For training the neural network, an additional step – extracting one of the folds of the cross-validation set to search for the optimum parameter configuration – was taken, a crucial step in order to get useful results in these types of statistical models (Zhang, 2007). Since the training is performed 100 times, a set of 100 results is obtained for both the parameters and the performances of the models, which allows to estimate a mean and a standard deviation for both of them. In all tables that are presented the notation "mean $\pm$ std. deviation" is used.

The results for the different models can be seen in Table 2.5. The results displayed are aggregated, that is, the obtained probabilities are transformed into Defaulters and Non-Defaulters, which is performed by simple addition of the probabilities of default. This is done in order to facilitate comparison with the binomial model, which differentiates between $\{P, (C \cup W)\}$. Theoretically, the combination could bring improvements in accuracy even if one score is sufficient – presents stochastic dominance – when compared to the other (Zhu et al., 2001), but since the models we use are complementary, we use the direct methodology for comparison. This comparison also takes into account the fact that the final goal is to have a better discrimination between these two classes, defaulters and non-defaulters. The results obtained are more clearly viewed when the Area Under the Curve is considered and the Receiver Operator Characteristics (ROC) curves are plotted.

Table 2.5: AUC for the three models.

| **Model** | **AUC (defaulters)** |
|---|---|
| Logistic Regression (w/o defaulter diff.) | 0.6275±0.017 |
| Multinomial Logistic Regression (w/ defaulter diff.) | 0.6678±0.004 |
| Neural Networks (w/ defaulter diff.) | 0.6660±0.004 |

Table 2.6: Parameters obtained from neural network training with defaulter differentiation.

| **Variable** | **P** | **C** | **W** |
|---|---|---|---|
| Activity_A | 0.06±0.01 | -0.29±0.03 | 0.16±0.02 |
| Activity_B | 0.05±0.01 | -0.16±0.02 | 0.09±0.01 |
| Owner | 0±0 | -0.1±0.62 | -0.07±0.01 |
| Tenant | -0.09±0.01 | 0.28±0.03 | -0.21±0.02 |
| Middle_Own | -0.09±0.01 | 0.21±0.03 | -0.12±0.02 |
| No_Property | -0.15±0.02 | 0.07±0.03 | 0.11±0.01 |
| One_Property | -0.05±0.01 | 0.01±0.01 | 0.04±0.01 |
| Region_1 | -0.17±0.02 | 0.16±0.03 | -0.03±0.01 |
| Region_2 | 0.12±0.02 | -0.21±0.03 | 0±0 |
| Guarantor | 0.06±0.01 | 0.09±0.01 | -0.15±0.02 |
| Term | -0.44±0.18 | -0.45±0.46 | 1.12±0.18 |
| Age | 7.31±1.13 | 3.91±1.88 | -11.99±1.93 |

Both are classical tests for measuring credit scoring quality (Brown and Mues, 2012), since they consider that the results are probabilities, and as such the threshold used to estimate the class of an element is an additional degree of freedom that can improve results. Table 2.5 shows the AUC that is obtained for each model, and Figure 2.1 displays the comparison between the models. The proposed methodology obtains a result that is from 5 to 7 percent better than the logistic regression model, with the AUC being statistically insignificant between the two proposed models, but both being superior to the logistic regression. In the ROC curve it also appears that the models are better at identifying defaulters in the upper range of thresholds, showing a significant difference from 0.3 onwards. This suggests the increase can be attributed to the improvement in discriminating between the two types of defaulters.

In Tables 2.6, 2.7 and 2.8 the parameters from the experiments can be observed. The first conclusion that can be drawn from the parameters is the low standard deviation that they possess, indicating that the solutions are very stable. The improvement in classification comes from those variables whose behaviour for the good payers lies in between the two different classes, as happens, for example, with the variable “Tenant”. This variable is not significant in the logistic regression

Table 2.7: Parameters obtained from multinomial logistic regression training with defaulter differentiation.

| **Variable** | **P** | **C** |
|---|---|---|
| Activity_A | 0.36±0.04 | -0.4±0.05 |
| Activity_B | -0.25±0.02 | -1.37±0.03 |
| Owner | -0.08±0.02 | -0.71±0.02 |
| Tenant | 0.18±0.02 | 0.55±0.03 |
| Middle_Own | 0.31±0.03 | 1.46±0.03 |
| No_Property | 0.03±0.04 | 1.01±0.04 |
| One_Property | -0.76±0.01 | -0.14±0.02 |
| Region_1 | -0.25±0.03 | -0.11±0.03 |
| Region_2 | -0.43±0.03 | 0.55±0.02 |
| Guarantor | 0.29±0.02 | -0.69±0.03 |
| Term | 0.62±0.02 | 0.72±0.03 |
| Age | -0.41±0.05 | -0.29±0.06 |

model. It is though in the other two models, with both models showing that being a tenant is a characteristic present for customers who do not have the capacity to repay, and that it is a neutral characteristic for good payers. Owning a house, in turn, is not a characteristic of customers with lack of willingness to pay. Another example is the variable "Guarantor". Having a guarantor means one is more likely to want to pay but fails to do so, than to be one who will repay, and that one is least likely not to want to pay. Of course the lender may create this by only requiring guarantors from borrowers who the lender thinks may not be able to repay.

Other interesting characteristics are that the customers in class *W* usually ask for longer terms, and adding this to the smaller amounts that they borrow implies an interesting behaviour: a customer who may default because they do not want to pay will borrow a small amount of money due at the end of a long term, which can be paid with small instalments, contrasted with borrowers of class *C* who are prone to a greater risk given that the instalments are larger. Additionally, good payers are associated with owning more than one property, or with no property at all (which is related to the segment from where the customers are extracted).

When analysing variable "Age" the multi-class models present contradictory information, with neural networks associating age with good payers, in lower relevance with defaulters of class *C*, and also indicating that customers in class *W* will be younger. This is inverted in both the logistic regression model and the multinomial logistic model, where greater age is associated with default-

Table 2.8: Parameters obtained from logistic regression training without defaulter differentiation.

| **Variable** | **Defaulter** |
|---|---|
| Activity_A | 0.33±0.04 |
| Activity_B | -0.29±0.02 |
| Owner | -0.27±0.02 |
| Tenant | 0±0.02 |
| Middle_Own | 0.39±0.03 |
| No_Property | 0.39±0.03 |
| One_Property | 0.69±0.02 |
| Region_1 | 0.2±0.02 |
| Region_2 | 0.73±0.03 |
| Guarantor | -0.56±0.03 |
| Term | -0.25±0.03 |
| Age | 0.19±0.04 |

ing. The reason for this might be that neural networks handle more complex patterns better, and age usually presents a non-linear behaviour in which borrowers present a larger relative risk of defaulting when they are young (usually 18-29 years old), but then present a decreasing risk at ages of 30 and up, until the risk rises again between the ages of 50-65. This behaviour differs from what is usually observed in US and UK databases, and can be explained by the increased expenses that these borrowers must face at these ages, when contrasted to their income. A common work-around is to segment variables into dummies, but one of the ideas of presenting the results in this way was to test the behaviour of the models analysing these contrasting patterns.

## 2.5 Conclusions and Future Work

In this paper, a methodology was presented for improving credit scoring results using constrained clustering arising from a model that describes the expected rational behaviour of lenders and borrowers. The new tools that have arisen in the last decade allow for a seamless and direct way to improve statistical models using different techniques, as we have done with economic behaviour and game theory. The combined applications of data mining, statistical analysis, and game theory are already being used in a large number of fields, and we present how they also show promise in financial analysis, where the behaviour of different agents is key.

The results presented show a significant improvement in creating scores over the common methodology, with a six percent average increase in discrimination capacity, which can be fairly significant when using this model considering the common amounts involved in consumer lending. However, an even greater contribution is extracted from the insights that are gained when studying defaulters. With the new goal of differentiating defaulters in mind, the results can be improved by determining variables that directly tackle the differences between the new classes. This is particularly relevant given that the improvements in the classification capacity can be associated directly with those variables where the behaviour of good payers lies in between the behaviour of the two other classes.

Also relevant are the insights that can be gained by studying the characteristics of the two new classes found. The variables that were used here brought very interesting information about how defaulters behave and can be detected, and that knowledge can be used to improve the basic requirements for loans, or to detect risk segments of the portfolios of loans in order to improve collecting campaigns. Better understanding of this segment also has the capacity to lower the overall risk present in the portfolios, with large potential savings.

The procedure in this work focused on applying semi-supervised techniques to identify unknown patterns, aided by a very simple decision process extracted from economic modelling. This research can be extended by designing more complex games that reflect the decision processes that defaulters make in greater detail, for example focusing on capturing irrational behaviours using behavioural economics. Some results in this setting can be seen for example in the work of Benton et al. (2007), among others. The second approach that can be followed is calculating a general equilibrium, eliminating the exogenous differentiation of defaulters and obtaining a characterization of the borrowers based on an endogenous decision process. In both cases, careful estimation of a set of constraints that would allow for knowledge discovery techniques to be applied would have to be performed, since a complex economic model could lead to equilibria that is not easily transformed to constraints as was the case in our work.

Finally, the main conclusion of this work is that traditional credit scoring can, and should, be improved by the use of more sophisticated techniques. The current developments in the fields of statistics, economic analysis, and behaviour, are powerful tools that are available for researchers and practitioners, who can benefit from improved modelling.

# Acknowledgments

The first author acknowledges the Chilean National Council for Research, Science and Technology (CONICYT) for the grants that support this work (AT-24110006), and to C. Mora and S. Beckman for their aid in editing this paper. This paper has been partially funded by the Institute of Complex Engineering Systems (ICM: P-05-004-F, CONICYT: FBO16) and the Finance Center of the Department of Industrial Engineering of the University of Chile.

# Chapter 3

# Semi-Supervised Constrained Clustering with Cluster Outlier Filtering (Unabridged)[1][2]

**Cristián Bravo and Richard Weber**

**Abstract**

Constrained clustering addresses the problem of creating minimum variance clusters with the added complexity that there is a set of constraints that must be fulfilled by the elements in the cluster. Research in this area has focused on "must-link" and "cannot-link" constraints, in which pairs of elements must be in the same or in different clusters, respectively. In this work we present a heuristic procedure to perform clustering in two classes when the restrictions affect all the elements of the two clusters in such a way that they depend on the elements present in the cluster. This problem is highly susceptible to outliers in each cluster (extreme values that create

[1]The following is an unabridged version of the book chapter published in "Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications". The original publication is available at www.springerlink.com, in particular http://www.springerlink.com/content/3410k18x2212t81m.

[2]Please cite this paper as follows: C. Bravo and R. Weber. Semi-supervised Constrained Clustering with Cluster Outlier Filtering. Lecture Notes in Computer Science 7042: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. César San Martín and Sang-Woon Kim, eds. Pages 347-354. 2011.

infeasible solutions), so the procedure eliminates elements with extreme values in both clusters, and achieves adequate performance measures at the same time. The experiments performed on a company database allow to discover a great deal of information, with results that are more readily interpretable when compared to classical k-means clustering.

## 3.1 Introduction

Among all methods for unsupervised pattern recognition, clustering (Xu and Wunsch, 2008) is possibly the most widely used technique. One of the research areas that has been receiving increasing attention in the last decade is the use of additional information regarding the problem, incorporated to the problem by using constraints that the elements must satisfy. This area is called "constrained clustering" (Basu et al., 2008), and has been applied to a wide range of fields (Bard and Jarrah, 2009; Levy and Sandler, 2008; Patil et al., 2006).

The methods of constrained clustering present a semi-supervised approach to obtain segments of a dataset incorporating certain restrictions that the members of one cluster, the members of different clusters, or the general structure of the clusters must fulfil. The term "semi-supervised" refers to the incorporation of knowledge that is not directly present in the data, or that is known only for a limited number of cases, in order to improve the results on the entire domain.

In this paper, a method to perform semi-supervised clustering with two classes is presented, that differs from the usual formulation in a key aspect: the constraints have to be satisfied by all the elements present in a cluster, against all the elements in the other cluster, clearly a challenge since simply changing one element from one cluster to another changes the whole set of constraints. Such application is common in social phenomena, where it is expected that the members of one group are, in some aspects, different than all the elements of the other cluster. One example is clustering customers whose client value is different, or segmenting groups with a defined utility.

In order to solve this problem, a heuristic method that takes into account the structure of the problem will be presented, to then test the proposed approach against classical k-means clustering, using a database of unreturned (defaulted) loans.

This work is structured as follows: Section 2 provides a brief overview of semi-supervised

clustering, to then present the proposed algorithm in section 3.3. Both a synthetic and a real-world application of the method and its results are presented in section 3.4, and finally conclusions are drawn in section 3.5.

## 3.2 Constrained Clustering and Semi-Supervised Methods

There are two different approaches reported in the literature for constrained clustering, as noted for example in Davidson and Ravi (2005), and both are based on the concept of "Must-Link" and "Cannot-Link" restrictions. The first set of constraints indicates that two elements must always belong to the same cluster, whereas the second one does not permit the presence of two elements in the same cluster.

The methods that can be constructed from these pairwise constraints have been studied in depth, and usually differ in the role of the respective constraints: in the first case the algorithm fulfils an objective or distance functions using the information from restrictions, the best known application of this work appearing in Basu et al. (2004), which uses Hidden Markov Random Fields to estimate the probability of belonging to each cluster.

In the second type of models, the constraints limit the presence of elements in one cluster or the other, using some heuristic approach to change and alter the clusters and converging to a new solution. One example of this type of procedures is the one proposed by Wagstaff et al. (2001) where the authors propose to alter the k-means procedure and iteratively construct clusters. A similar approach is followed in this paper adapting it to the new problem: each of the elements of one cluster must satisfy the restrictions against most of the elements of the other cluster. The inclusion of these constraints for each instance has not received such extensive study so far, with some attempts to select only relevant constraints to, for example, make the problem more tractable and select the information that is more relevant (Zhao et al., 2011).

There are several issues that must be addressed in the problem stated above. First, each time an element is changed from a cluster many restrictions have to be recalculated and re-checked. Second, the high number of restrictions can make the problem intractable. And finally, if there is an element with an extreme value in one of the clusters, that is, an element whose variables

make the value of the constraints too high or too low, the problem can easily become infeasible. To approach this, in the next section we will present a heuristic approach to solve the problem efficiently.

## 3.3 Iterative Procedure for Constrained Clustering with Extreme Value Elimination

The procedure presented here differs from previous works in that it filters outliers in each iteration, checking for the best solution in terms of violations and cluster robustness. In each iteration the results are not necessarily feasible, because such search is an NP-hard problem and it would make the algorithm impractical (Davidson and Ravi, 2005). With this modification, the procedure is very fast, as it is necessary when clustering medium to large datasets.

If we assume a dataset $X$ with $N$ elements, and a partition of the dataset into two clusters, such that each one ($X_1$ and $X_2$) is formed by elements $x_i^1 \in X_1$, $i \in \{1,\ldots,I_1\}$ and $x_j^2 \in X_2$, $j \in \{1,\ldots,I_2\}$, with centroids $c_1$ and $c_2$, the problem is to solve the following optimization problem, adapted from a common formulation by Doğan and Güzeliş (2006):

$$
\begin{aligned}
\min_{M,c_1,c_2,\mathrm{ext}_1,\mathrm{ext}_2} \quad & \sum_{i=1}^{N}\sum_{k=1}^{2} \|x_i - c_k\|^2 \cdot m_i \\
s.t. \quad & M \in \{0,1\}^N \\
& R_1 \cdot x_i^1 \geq \mathrm{ext}_1 \quad \forall x_i^1 \mid m_i = 1 \\
& R_2 \cdot x_j^2 \geq \mathrm{ext}_2 \quad \forall x_j^2 \mid m_i = 0 \\
& \mathrm{ext}_1 \geq b_2 \cdot x_i^2 \quad \forall x_i^1 \mid m_i = 0 \\
& \mathrm{ext}_2 \geq b_1 \cdot x_j^1 \quad \forall x_j^2 \mid m_i = 1
\end{aligned}
\tag{3.1}
$$

Where $R_1$ and $R_2$ are the set of parameters associated with the constraints that each case must satisfy, against all the other functions of the values in the other cluster, represented by vectors $b_1$ and $b_2$ and extreme values $\mathrm{ext}_1$ and $\mathrm{ext}_2$. Vector $M$ indicates the cluster to which element $i$ belongs,

taking a value of 1 if the element belongs to cluster 1 and of 0 if it is not. This formulation is a reduced form used to illustrate the complexity of the problem, since the variable $m_i$ is present in the definition of the restrictions. A more explicit version would have $N \times N$ restrictions, one for each element and cluster, since *a priori* it is not known in which cluster each element is, nor the centroids of each cluster.

In order to solve this problem, we propose a heuristic procedure that takes into account the constraints that must be satisfied and takes advantage of the fact that the linear restrictions are bound to possess a maximum or minimum value in the dataset, that is, that given a fixed distribution of elements in the cluster, only the values $ext_1$ and $ext_2$ have to be checked against the elements of the cluster. The procedure starts with random centroids, and in each iteration the elements in the cluster are compared to the extreme value of the other cluster, i.e., if the element is in cluster 1, then it is checked whether restriction $R_1 \cdot x_i^1 \geq \text{ext}_1$ is satisfied with the largest or smallest element present in cluster 2, according to what is necessary. In case at the end of the movements the conditions are not fulfilled by all cases, then the extreme values in both clusters are removed from the analysis and the process is repeated. The algorithm continues until both the violations are below a threshold and the values of the centroids do not move more than a given tolerance. The algorithm from constrained clustering with filtering (CCF) is described in Algorithm 1.

The elements in each cluster have to satisfy the restrictions against most of the elements in the other cluster, and this is accomplished by eliminating a small number of extreme cases in each iteration and seeking that all other satisfy the constraints. At each step the minimum variance cluster is approximated (by assigning elements to the closest cluster), and is this solution the one that is perturbed by moving the elements according to the constraints. The convergence of the algorithm is ensured, since at worst case (infeasible problem) only two elements will remain and the method will stop with one element per cluster.

The algorithm was implemented and is available as a free R package "multiscore"[3] along with all the rest of the code used for this PhD thesis.

[3]The package is available at `http://dmgroup.cf.dii.uchile.cl/descargas/`

**Algorithm 1** CCF(Dataset $X$, $R_1$, $R_2$, $b_1$, $b_2$)

$C = (C_1, C_2) \leftarrow$ `Random(size(2))`
`Flag` $\leftarrow 1^N$ {*If element is used or outlier*}
**while** Movement in $C > \varepsilon$ **do**
    Assign elements to closest cluster
    $M1 \leftarrow X($`cluster` $= 1)$
    $M2 \leftarrow X($`cluster` $= 2)$
    Calculate $\text{ext}_1$ and $\text{ext}_2$ from vectors $M2 \cdot b_2$ and $M1 \cdot b_2$
    Violations $\leftarrow 0$, Eliminated $\leftarrow 0$
    **while** Eliminated $< 0.01N$ **or** Violations $> \varepsilon N$ **do**
        **for** $i = 1$ **to** $N$ **do**
            **if** `Flag`$(i) = 0$ **then**
                Skip $i$
            **end if**
            **if** `Cluster`$(x_i) = 1$ **then**
                $R \leftarrow R_1 \cdot x_i$
            **else**
                $R \leftarrow R_2 \cdot x_i$
            **end if**
            **if** Element $i$ violates conditions **then**
                Change cluster of element
                `Violations` $\leftarrow$ `Violations` $+ 1$
            **end if**
        **end for**
        **if** `Violations` $> \varepsilon$ **then**
            `Flag`$(I(\text{ext}_2)) \leftarrow 0$, `Flag`$(I(\text{ext}_1)) \leftarrow 0$
            `Eliminated` $\leftarrow$ `Eliminated` $+ 2$
        **end if**
        `Recalculate`$(M1, M2, \text{ext}_1, \text{ext}_2)$
    **end while**
    `Recalculate(C)`
**end while**

## 3.4 Experimental Results

To test the algorithm, two datasets were used. The first is a toy dataset from the multiscore package, which is described in the next subsection. The second is a real dataset from credit scoring[4], as will be explained in section 3.4.2.

[4]N.R: The same dataset from Chapter 1.

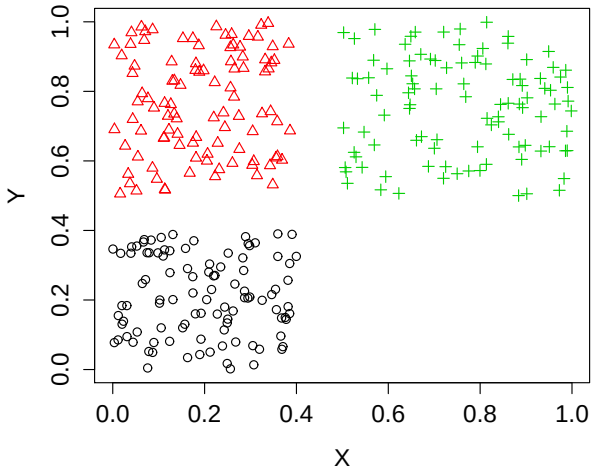



Figure 3.1: Original dataset, with three clusters and clear low density zones.

### 3.4.1 Toy Data

The first experiment uses toy data shown in figure 3.1. The figure depicts 300 points spread across three segments (clusters) of the $[0,1]^2$ region, with clear low-density zones. The decision to be made is what happens with the cluster in the upper-left zone, when clustering is performed using just two groups. Figure 3.2 shows the result of classical K-Means clustering: the upper-left segment is spread evenly among the two groups, which does not seem sensible.

One approach, for whom CCF algorithm can be applied, is limit a priori the cases that might enter the second cluster. In particular, one could limit the first cluster such that coordinate $X$ for the elements of the first cluster must never be greater than the coordinate $X$ of the elements of the second cluster, that is: $R_1 = (1,0)$, $ext_1 = \min\{(1,0) \cdot V : V \in D, m_1 = 0\}$, $R_2 = 0$, $\wedge ext_2 = 0$, with $D$ the dataset. The results can be seen in Figure 3.3.

The effects of the algorithm are clear. The upper-left cluster is assigned, as it should be, to the first cluster. Furthermore, convergence is quickly achieved (four iterations), the first one assigning most of the cases correctly, and the rest simply moving the centroids to the most appropriate positions following the K-Means procedure. No cases were eliminated, since in each iteration all constraints were satisfied. Finally, convergence time is very fast, just a few seconds in an cc2.8xlarge instance from the Amazon Elastic Compute Cloud (EC2) service.

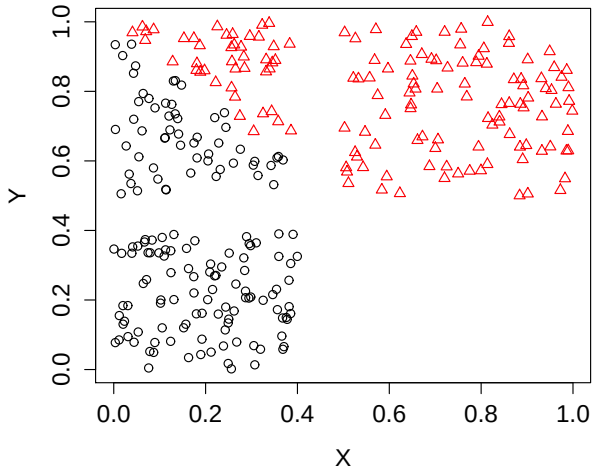



Figure 3.2: Result of applying K-Means clustering with $K = 2$. The upper-left cluster is split in two.

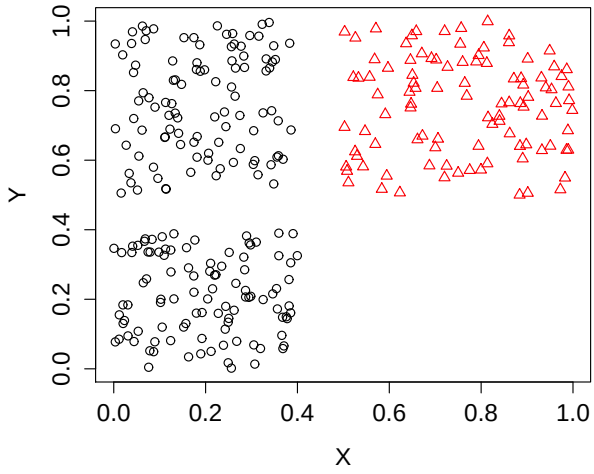



Figure 3.3: When the CCF algorithm is applied to the dataset, the upper left-cluster is assigned to cluster 1 in its entirety, as expected.

### 3.4.2 Real Data

To test the presented methodology, a dataset consisting of 24,500 loans granted to mass-market is available, all of which were not returned. The database originates from a Chilean organization, and comprises a 10 year period, from 1997 to 2007. Each loan is described by the following variables, which are associated to the customer or to the loan itself:

- Collaterals: The collaterals are described by two variables. The first is a dummy variable that indicates whether the customer secured the loan or not (With_Collaterals), The second one represents the value of the collateral (Value_Collateral_UF), in UF, the Chilean inflation-adjusted monetary unit.
- Amount and Rate: The amount of the loan, in UF, and the total annual interest rate charged for the loan.
- Arrears and Cancellations: The total sum of days the instalments of the loan were in arrears before defaulting is included in variable Days_Arrear. Also, sometimes the institution will cancel the payment of punishments and excess interest that arises from arrears. This event is resumed into two different variables, considering the number of times this happened in the loan lifetime (Num_Cond), and the amount that was reduced (Amount_Cond). Additionally if some of the interests due to be paid are also discounted from the instalments, this value is annotated in variable Interest_Low.
- Extensions: Sometimes the company will extend the period of an instalment. The number of times a customer applies to this benefit appears in variable Num_Post, and the amount adjusted appears in Amount_Adjust, and, since the adjustment can be positive or negative, the total amount of negative adjustments is incorporated into Negative_Adj.

### 3.4.3 Constraints

In this case, a set of constraints is created from the economical behaviour expected between the customers, as shown in Chapter 1. In particular, the set $R$ of restrictions characterizes the rationality of not paying back the loan for two reasons: failure in capacity of repayment (class $G$), and failure in willingness of repayment (class $W$). In this setting, customers request an amount $x_G$ or $x_W$ of money, are charged a rate $r$, and discount their income using discount rates of $\delta_W$ and $\delta_G$. The lender has requested a collateral valued at $C_i$ for each customer $i$, and discounted by the company by a value of 40%, that is $\alpha = 0.6$, which is a known value, and discounts its income by a value of $\delta_C = 0.93$. There is an external chance of losing income (which translate into failure in paying back the loan) given by $q = 0.15$, extracted from the long term default rate of the institution, and an *a priori* belief that the customer is of class $W$ given by $\theta = 0.55$, arising from internal estimations of the company, as well as parameters $\delta_G$ and $\delta_W$ which were fixed to values of 0.75 and 0.5

respectively. Under this setting, and assuming linear utilities, it is possible to estimate the amounts that should be requested such that rationality is achieved. The restrictions that are proposed are:

$$\begin{aligned} x_{G1}(1-\delta_G(1-q)(1+r))+\delta_G x_{G2}(1-q)\cdot \\ (1-\delta_G(1-q)(1+r))-\delta_G qC_1-\delta_G^2(1-q)qC_2\geq 0 \end{aligned} \tag{3.2}$$

$$\begin{aligned} \theta x_{G1}(-1+\delta_C(1-q)(1+r))-(1-\theta)x_{W1}+x_{G2}\delta_C\theta\cdot \\ (-1+\delta_C(1-q)(1+r))+C_1\alpha\delta_C(1-\theta(1-q))+\theta(1-q)q\delta_C^2\alpha C_2\geq 0 \end{aligned} \tag{3.3}$$

$$x_{W1}\geq C_1\delta_W q \tag{3.4}$$

These restrictions come from assuming that customers in class $G$ desire a second loan, while customers in class $W$ do not. The extensive formulation of this game-theory problem is not relevant to the clustering procedure itself, so it will be omitted in this work.

### 3.4.4 Results

The experiments are run on the normalized database, and then the results are de-normalized to better reflect the differences obtained, the method was implemented using R[5]. The method eliminates only 2% of the total cases, and converges in two minutes which, given that the database is medium sized, it is a very good convergence time. To test the stability of the method, the procedure was run 10 times from different random starting points, all converging to roughly the same result as is to be expected.

To study the information that the model brings the centroids of the cluster must be studied. Table 3.1 (left) presents the obtained results. From the clustering procedure it arises that the differences in collaterals are important but have to be interpreted carefully, since the collateral value is used in the constraints. More relevant is that the percentage of customers with a collateral is not really meaningful between the clusters, so the conclusion is that is not the presence of a collateral the relevant information, but its value. Also, customers in class $G$ request a far larger amount for

[5]In the original publication, the code was implemented in MATLAB. This has been changed in this version

Table 3.1: De-normalized centroids for semi-supervised clustering procedure (left) and k-means procedure (right).

(a) Semi-Supervised Algorithm

| **Variable** | **Class *G*** | **Class *W*** |
|---|---|---|
| With_Collaterals | 0.41 | 0.33 |
| Amount | 52.94 | 14.82 |
| Days_Arrear | 895.78 | 719.82 |
| Num_Cond | 0.27 | 0.36 |
| Num_Post | 0.85 | 0.46 |
| Num_Reneg | 0.72 | 0.45 |
| Amount_Adjust | -0.27 | -0.08 |
| Amount_Cond | 4.23 | 2.44 |
| Negative_Adj | 0.41 | 0.09 |
| Interest_Low | 1.32 | 0.31 |
| Payments | 2.81 | 2.42 |
| Value_Coll_UF | 71.37 | 7.93 |
| Rate | 1.10 | 1.10 |

(b) K-Means Clustering

| **Variable** | **Cluster 1** | **Cluster 2** |
|---|---|---|
| With_Collaterals | 0.35 | 0.34 |
| Amount | 21.78 | 22.57 |
| Days_Arrear | 1242.10 | 337.93 |
| Num_Cond | 0.56 | 0.15 |
| Num_Post | 0.42 | 0.64 |
| Num_Reneg | 0.68 | 0.35 |
| Amount_Adjunt | -0.15 | -0.05 |
| Amount_Cond | 4.63 | 1.21 |
| Negative_Adj | 0.26 | 0.06 |
| Interest_Low | 1.06 | 0.04 |
| Payments | 2.25 | 2.70 |
| Value_Coll_UF | 18.55 | 21.64 |
| Rate | 1.10 | 1.10 |

their loans, which would be consistent with a default based on the capacity of payment. Considering the total number of days in arrears, class *G* accumulates 170+ days more than class *W* before defaulting, indicating that they make a greater effort of paying back the loan than class *W*. The procedure also shows that they apply for a larger number of renegotiations (0.74 per customer in average), get greater adjustments and debt relieves, and are more prone to receive a discount on their due interests (1.32 UF per customer in average, versus 0.31). Finally, they pay almost one full instalment more than the customers in class *W*.

The values of the variables hint that there is indeed a different behaviour detected, but to shed light on whether this procedure brings indeed more information than classic K-Means clustering, Table 3.1 (right) presents the results of a K-Means clustering procedure that will be used for comparison. Stability comparisons, such as Davis-Bouldin index, are not relevant, since obviously the constrained algorithm will have a greater standard deviation than the unconstrained problem.

The two tables show the advantages of incorporating additional information in the form of constraints, since the k-means procedure focuses on two variables: days in arrears and reduction in rate (Interest_Low). All the other variables present only minor differences or equivalent results to the proposed method. The conclusion that can be extracted from the k-means procedure is that there are indeed two groups with some difference in their payments behaviour, but the information that

can be deduced does not permit a more meaningful interpretation. This is contrasted with the much more rich information that the semi-supervised clustering procedure brings, where the differences between this variables are also present, but are also complemented with a series of other, more subtle, differences that arise from the restrictions imposed.

## 3.5 Conclusions

A procedure to estimate the centroids of a two-class constrained clustering problem was presented. The main difference between this procedure and the ones presented previously in the literature is that the constraints are associated to all elements in each cluster, as well as allowing intra-cluster restrictions. This problem is much harder to solve than traditional constrained clustering with must-link and cannot-link constraints, since each time an object changes its cluster, all the constraints have to be re-checked. Additionally, since all the objects must be checked with respect to all restrictions, it is usually not possible to satisfy the constraints for all elements.

The proposed method is a heuristic procedure based on first obtaining minimum variance clusters and then adjusting the constraints and filtering outliers. The experimental results show that it performs fast and reliably, presents few eliminated cases, and shows a reasonable convergence time.

However, the most important feature of the method is that the results use correctly the additional information that is considered through the constraints. When applying the method to a database of defaulted loans, the results from a classical K-Means algorithm are greatly enriched, presenting more subtle differences between the classes and profiling two different segments of defaulters using only expected rational behaviour.

It can be concluded that the model is useful for the presented problem, and that the results show the benefits of including external information in clustering procedures. Future work in this line is to use this results to improve classification in credit risk problems.

# Acknowledgments

The first author would like to acknowledge CONICYT for the grants that finance this research, to the Ph.D. program in Engineering Systems for their support in the development of this work, and to C. Mora for her aid in editing this paper. Also, support from the Institute on Complex Engineering Systems (ICM: P-05-004-F, CONICYT: FBO16) (www.sistemasdeingenieria.cl) is acknowledged.

# Chapter 4

# Survival Analysis for Credit Scoring with Multiple Types of Defaulters[1]

**Cristián Bravo, Lyn C. Thomas, and Richard Weber**

**Abstract**

Differing from traditional models that approximate whether an event occurs or not, survival analysis is a statistical technique which measures the time until an event occurs. We study an extension of survival analysis that deals with events that may occur due to more than one reason, contrasting two approaches for multi-class survival analysis: competing risks and mixture models, which differ both in functional form and in the basic assumption on how the events occur. The contrast is performed in a dataset which takes explicitly into consideration whether default occurs due to problems of ability to repay, or because of willingness to repay. We benchmark both multi-class methods against binary logistic regression and Cox's proportional hazard binary regression, using a database of 5,000 loans to low-to-middle income micro-entrepreneurs. The results show that even though mixture models estimate a more complex function and define more detailed profiles of default, there is not a significant gain in predictive capability when contrasted to competing risks, with both models very close to each other in terms of predictive capability. Both multi-class models, however, show a small gain in predictive capability

[1]The following is an up-to-date (to the time of writing) version of the paper "Survival Analysis for Credit Scoring with Multiple Types of Defaulters", please do not cite this paper without authorization.

when compared to both logistic regression and binary survival analysis. The most notable gains come, however, from obtaining much more detailed predictive variables for each of the classes, since the mixture model delivers different variables for each class and each survival function, which in turn leads to a better understanding of default.

**Keywords:** Credit Scoring, Survival Analysis, Statistics, Domain-Knowledge, Constrained Clustering, Banking.

## 4.1 Introduction

In credit scoring the objective is to create a measure of the risk a lender is undertaking when borrowing a certain amount of money to an applicant (Thomas et al., 2002). There are two approaches to model this problem: by modelling whether the event occurs in a given time period, the dichotomous approach, or by modelling the time-to-event, the survival approach. In the last few years, an increasing amount of attention has been given to the survival method, with several publications presenting the advantages of using this different paradigm.

The use of survival analysis in credit scoring starts with Narain et al. (1992), and its use has been analysed more deeply by Banasik et al. (1999) and Stepanova and Thomas (2002). In all those works, the importance of having a hazard rate of default that depends on the age of the loan is is emphasized, which is stated as a support for provisioning, and for better comprehension of the event of defaulting. Since then, it has been applied in several fields and with several variations: Baesens et al. (2005) used neural networks to construct survival models, Bellotti and Crook (2008) studied the impact of macroeconomic variables in credit scoring using survival models, Malik and Thomas (2009) used survival models to study the relationship between corporate bonds and consumer credit risk models, Glennon and Nigro (2005) studied default risk for small businesses, Allen and Rose (2006) applied survival analysis in the context of loans in New Zealand, and, more recently, Tong et al. (2012) applied mixture cure models – a variant of survival analysis that incorporates a class of non-defaulters which would be defaulters given enough time – in credit scoring.

In this work we extend the survival modelling for credit scoring approach further, analysing what happens when there are more than two classes of defaulters. In this case, classical survival

models must be extended to include that the default event can happen because of different reasons. In the context of credit scoring, the use of these models has only been hinted at in other works (Stepanova and Thomas, 2002), in which finishing the loan due to prepayment or due to defaulting was the multi-class event. We consider using our previous work (Bravo et al., 2013) where default occurs because of two well-known causes: the inability to repay, and the unwillingness to repay. Our paper focuses on survival models with different paradigms and assumptions, and questions the appropriateness for credit scoring under multiple classes.

Our work uses survival analysis in credit scoring, both in binary and in multi-class methods, studying their properties; it illustrates the use of economic modelling to differentiate reasons for default; and shows the effects of differentiating defaulters and using multi-class survival models in the context of credit scoring. We also present the results of applying the methodology on a database of loans granted to low-to-middle income micro-entrepreneurs from a Chilean private financial institution.

This paper is structured as follows: The next section presents the model for using of binary survival modelling in credit scoring, followed by the extension to multi-class problems. The next section presents the methodology to construct the multi-class dataset, followed by the experimental results of applying the procedures to the dataset mentioned above. We place special emphasis on the knowledge that can be extracted by using the techniques. In the final section, conclusions are drawn and future work is presented.

## 4.2 Overview of Survival Analysis in Credit Scoring

Survival analysis techniques (Hosmer et al., 2008) are a set of statistical techniques which model the time-to-event using available data. Having being developed for the biological sciences, it has been proved useful in many other areas. It is suitable for use in credit scoring since there is a process (the repayment of loans) that could fail in any instalment of the loan.

A lender has information on a set consisting of $N$ borrowers, each of whom has had a loan granted. This information is described in a set of $v_p$ characteristics (regressors) $X_n^p \in R^{v_p}$, for each borrower $n \in \{1,\ldots,N\}$, and information on the performance of the loan, in a set $X_n^f \in R^{v_f}$.

The first set $X_n^p$ includes socio-demographic information, and details of the loan application. The information set $X_n^f$ includes if and when the default event occurred. For the standard approach, this is described by $d_n \in \{0,1\}$, describing if the borrower paid back all the instalments of the loan in the first 12 months, but for the survival approach we require two other variables, $y_n \in \{0,1\}$ which describes if the borrower defaulted (1) or not (0) along all the instalments of the loan, and a time $t_n$, which is either the time of default or the time when no further information is available. This set also includes all the information that was collected post-granting: information regarding days in arrears, variables of possible renegotiation or condoning of debts, and any other behaviour-related variable that describes the evolution of the loan.

In the proportional hazards model, Cox (1972) proposes that the *hazard function* $h(t,x)$, the probability that a loan will fail in time $t$ given that the loan has not failed already is given by:

$$h(t,x_n) = h_0(t) \cdot e^{\beta \cdot x_n} \tag{4.1}$$

for each borrower $n$. $e^{\beta \cdot x}$ is a time-independent function of regressors ($x_n$) and parameters ($\beta$), and $h_0(t)$ or *baseline hazard* is a regressor-independent function that models the population hazard at each time. Since the hazards ratio – the ratio of the probability of default between any two subjects that have not yet defaulted at a fixed time – are independent of the time and proportional, we get the assumption of proportional hazards which names the model.

Once the hazards are defined, it is possible to determine the *survival function*, the probability that the time a borrower will survive $T$ is greater than $t$, given by:

$$\begin{aligned} p(T > t) &= e^{-\int_0^t h(t',x)dt'} \\ &= \left[e^{-\int_0^t h_0(t')dt'}\right]^{e^{\beta \cdot x}} \\ &= [S_0(t)]^{e^{\beta \cdot x}} \end{aligned} \tag{4.2}$$

in which $S_0(t)$ is the baseline survival function that depends only on time and not on the characteristics of the individual, and $\beta$ is the vector associated with parameters $x$. A partial likelihood is used

to estimate the $\beta$ parameters, that resembles the equations for logistic regression. The functional form of the likelihood corresponds to:

$$\mathcal{L}_p = \prod_{n=1}^{N} \left[ \frac{e^{\beta \cdot x_n}}{\sum_{i \in R_{t(n)}} e^{\beta \cdot x_i}} \right] \tag{4.3}$$

where $R_{t(n)}$ represents the set of all borrowers that have survived up to the same time as borrower $n$, $t(n)$. To maximize this function several approximations are implemented in most statistical packages. The two most common ones have been presented by Breslow (1974) and Efron (1977).

Finally, to obtain the baseline hazard function, an extension of the procedure by Kaplan and Meier (1958) is used, which builds the baseline survival function by generating estimates of the proportion of cases that still have not defaulted after time $t$ and the elements that did default.

## 4.3 Survival Analysis with Multiple Classes

Now let us consider that there are $J$ classes of the default event (or of borrowers), with $J > 1$, and we would like to calculate the probability that an individual does not survive up to time $T$, i.e. $P(t < T)$. There are several approaches that can be followed to solve the problem, but we will present two fundamental approaches that are relevant for this case: the straight-forward approach of competing risks, and mixture models.

These two approaches differ on the basic assumptions that form the survival function and the hazard rates function. In the basic competing risk approach, there are several causes for occurrence of the studied event, which are represented as different hazard functions with varying parameters and baseline hazards. In the mixture cure model, the survival function is considered to be a total probability construct built from several survival functions which are linked by a response probability $\pi_j$. In the following sections both approaches are explained in detail.

### 4.3.1 Competing Risks

In the competing risk approach, one can assume that the events of failure due to the different classes are independent (Crowder, 2001), since one cannot distinguish empirically between an independent and a dependent model given certain hazard rate and survival functions. We define now $S_j(t,x)$, the survival function of class $j$, as the probability that an individual with characteristics $x$ has not suffered a class $j$ failure by time $t$. We also define the hazard function $h_j(t,x)$ by $h_j(t,x) = -\frac{d}{dt}\log(S_j(t,x))$ in the continuous time case, while $S(t,x)$, the probability that an individual with characteristics $x$ will have not defaulted by time $t$ is:

$$S(t,x) = \prod_{j=1}^{J} S_j(t,x) \tag{4.4}$$

This approach has been used in medical studies and other sciences for several applications (see for example Hosmer et al. (2008), Kalbfleisch and Prentice (2002), or Marubini and Valsecchi (2004), among many others) and in the context of credit scoring has been briefly analysed in Stepanova and Thomas (2002) and in Banasik et al. (1999). From (4.4) and the relationship between hazard and survival functions it follows that:

$$h(t,x) = \sum_j h_j(t,x) = \sum_j h_{0,j}(t)e^{\beta_j \cdot x} \tag{4.5}$$

that is, an individual is simultaneously at risk of all causes of failure. For every $j \in \{1,\ldots,J\}$ it follows that there is a probability function $S_j(t,x)$ which corresponds to the probability of surviving past time $t$ and being of class $C = j$:

$$P(T > t, C = j|x) = S_j(t,x) = (S_{0,j}(t))^{e^{\beta_j \cdot x}} \tag{4.6}$$

By construction, the unconditional probability of surviving past time $t$ corresponds to:

$$\begin{aligned} S(t,x) &= e^{-\int_0^t h(t,x)dt} = e^{-\sum_j \int_0^t h_{0,j}(t)dt \cdot e^{\beta_j \cdot x}} \\ &= \prod_{j=1}^{J} e^{-\int_0^t h_{0,j}(t)dt \cdot e^{\beta_j \cdot x}} \\ &= \prod_{j=1}^{J} \left( \underbrace{e^{-\int_0^t h_{0,j}(t)dt}}_{S_{0,j}(t)} \right)^{e^{\beta_j \cdot x}} \\ &= \prod_{j=1}^{J} \left(S_{0,j}(t)\right)^{e^{\beta_j \cdot x}} = \prod_{j=1}^{J} S_j(t) \end{aligned} \tag{4.7}$$

In order to standardize this result, the unconditional probability of surviving less time than $t$ – i.e. defaulting – corresponds to:

$$P(T < t) = 1 - S(t,x) = 1 - \prod_{j=1}^{J} S_j(t,x) \tag{4.8}$$

### 4.3.2 Mixture Models

A mixture model is commonly referred to as any combination of different models that are used to better approximate an event. For survival analysis an extension to the competing risks approach that used such models was presented by Larson and Dinse (1985). In this case, we assume there is a response function $\pi_j(z)$ that models the overall probability of failing due to cause $j \in 1,\ldots,J$, such that $\sum_j \pi_j(z) = 1$. The vector of characteristics $z$ corresponds to a set of variables which are regressors for the occurrence of event due to cause $j$, which can be different from the vector of characteristics $x$ used to describe function $h(t,x)$ (equation 4.1). Several approaches have been studied regarding the particular specification that the survival function must posses, but we will focus on the one which uses both the Cox functional form for the hazard rate function, and multinomial logistic regression for the response probability $\pi(z)$, as proposed by Ng and McLachlan (2003) and Escarela and Bowater (2008).

The main reason to use that particular specification is that the model will then maintain the basic

linearity properties that are necessary for credit scoring Siddiqi (2006), otherwise the model would be inapplicable in a corporate environment. Also, the statistical properties of the mixture model with logistic form has been studied in detail by Escarela and Bowater (2008), with the conclusion that it has well-behaved asymptotic properties, such as asymptotically normal parameters. The model is based on a survival function of the following form:

$$\begin{aligned} S(t,x,z) &= \sum_{j=1}^{J} \pi_j(z) \left[S_{0j}(t)\right]^{\exp(\beta_j \cdot x)} \\ &= \sum_{j=1}^{J} \frac{\exp(\gamma_j \cdot z)}{\sum_{k=1}^{J} \exp(\gamma_k \cdot z)} \left[S_{0j}(t)\right]^{\exp(\beta_j \cdot x)} \end{aligned} \tag{4.9}$$

The main theoretical difference between the competing risk approach and the mixture model presented above is explicitly defined in the previous equation. In competing risks, the hazard rates functions, and subsequently the survival functions, assume that hazard rates are additive so each element is at risk of all causes of failure at the same time. The occurrence of the event then happens when one risk cause manifests. In the mixture approach, an element is subject to only one source of risk at the same time, but to the observer it is unknown which cause of risk is the one the element is really subject to, so the observer assumes a probability distribution for the causes of risk. This fundamental difference is the one that should be taken into account when applying multi-class survival analysis: is the subject simultaneously at risk of failing due to all classes or is he subject to just one, but the cause is unknown? If the case is the former, then competing risks might be the more appropriate condition, but if it is the latter, then mixture models might be a better fit.

The functional form presented here is very similar to the mixture cure model (Peng, 2003), with the difference being that in the latter, the uncensored events are considered to be either long term survivors (the event was not present) or individuals which would have presented the event had they been given enough time. Mixture cure models are then particular cases of the more general model studied in this paper. An application of mixture cure models to credit scoring is presented by Tong et al. (2012).

To estimate the parameters of the functions, it is possible to construct and train a log-likelihood function from expression (4.9), which then can be solved by several methods. Kuk and Chen (1992)

uses an Expectation–Maximization (EM) procedure combined with Montecarlo simulations, but both Ng and McLachlan (2003) and Escarela and Bowater (2008) present a general methodology based only on a EM algorithm to obtain the parameters, which will be the one used in this work. The procedures are based on maximizing the complete data log-likelihood function given by:

$$\begin{aligned}\log(\mathbb{L}(\gamma,\beta)) = \sum_{i=1}^{n}\sum_{j=1}^{J}\Big[1_{(C_i=j)}\log(\pi_j(z_i;\gamma_j)h_j(t_i,x_i;\beta_j)S_j(t_i,x_i;\beta_j)) \\ +1_{(C_j=0)}w_{ij}\log(\pi_j(z_i;\gamma_j)S_j(t_i,x_i;\beta_j))\Big]\end{aligned} \tag{4.10}$$

where $1_{(C_i=j)}$ is equal to 1 when element $i$ is in class $j$, and $w_{ij}$ is a binary random variable for uncensored cases that represents whether case $i$ should have belonged to any of the defaulter classes given enough time. To maximize this function, the EM procedure iteratively solves the problem by using $\pi$ as an expectation of the value of $z$, calculating the accumulated hazard function for each class $H_{0j}$ using a Kaplan-Meier based estimator, and finally solving the conditional expectation resulting from replacing these values in the log-likelihood function.

In previous work, either the model had been presented theoretically without any further application, or the paper had been applied in simplified form ($x = z$) in real databases. In this paper we tackle both problems, and we have developed the R package "multiscore"[2] to solve them. The package allows for different sets for the survival and the logistic class and is publicly available.

## 4.4 Different Classes of Defaulters

In order for the multi-class models to be of any use, it is necessary to create a dataset which possesses more than two classes in the context of credit scoring. The model by Bravo et al. (2013) allows one to construct a dataset by further splitting defaulters into two classes: those who default due to problems with their capacity to repay, and those who are not willing to repay.

[2]The package is available for download at `http://dmgroup.cf.dii.uchile.cl/descargas/multiscore_0.10.2.tar.gz`.

The procedure was developed for consumer loans in portfolios that included both secured and unsecured loans. The methodology presents a set of economic constraints for borrowers and lenders so that the loan is profitable for both.

To construct the constraints, we assume that there are two types of borrowers: class $C$, customers willing to repay who may be subject to an external shock to their income which occurs with probability $q$, and class $W$, which are customers who do not intent to repay the loan when given to them. The customers discounts their income with discount factors $\delta_C$ and $\delta_W$, respectively, and also the lender possesses a discount factor $\delta_L$. The borrowers request an amount $a_C$ or $a_W$, according to their class, and if they repay the loan they may request a further amount $f_a \cdot a_C$ in the future, with $f_a > 1$. The lender may in turn request a collateral of value $Co$ for the first loan, and expects to be able to request a collateral $f_{Co} \cdot Co$ for the second loan, in case the borrower does not default. Finally, the lender believes that a fraction $\theta$ of borrowers are of class $C$, and if there is default it can recover the $\alpha \cdot Co$, with $1-\alpha$ the haircut on its value.

Under this setting, loans are granted and accepted if the following conditions are fulfilled for all pairs $(a_c, Co_c), (a_w, Co_w)$, such that $c$ is an element in the cluster for class $C$ and $w$ is an element in the cluster for class $W$ (Bravo et al., 2013):

$$\begin{aligned} a_w &\geq \delta_W Co_w \\ a_c(1-\delta_C(1-q)(1+r))(1+\delta_C f_a(1-q)) &\geq Co_c \delta_C q(1+\delta_C f_C) \\ a_c\theta(\delta_L(1-q)(1+r)-1)(1+f_a\delta_L(1-q))+ & \\ Co_c\alpha\delta_L(1-\theta(1-q)q\delta_L f_{Co}) &\geq (1-\theta)a_w \end{aligned} \tag{4.11}$$

To construct the two classes of defaulters, the idea is to split defaulters into two sets such that in each set defaulters satisfy the constraints given (4.11), using only variables that describe the relevant repayment history of the loan (the $X^f$ set of variables described before). The methods of constrained clustering (Basu et al., 2008), and in particular the CCF procedure by Bravo and Weber (2011) allows to find such sets. The methodology is an extension to well-known k-means clustering that in each iteration adjust the minimum variance solution to satisfy the constraints, and in case of failing, eliminating infeasible cases.

## 4.5 Experimental Results

The proposed approaches were tested in a database of loans granted to low-to-middle income micro-entrepreneurs during the years 2000-2007. We begin by describing the dataset and the variables that are available, and then present the construction of the new objective variable using economical differentiation. Finally, the results for the classification are shown.

### 4.5.1 Dataset Description

The database consisted of approximately 4,600 new applicants, with a default rate of 25 percent (1,111 applicants). The original dataset counted more than 400 variables describing socio-economical profile, income, expenses, debt structure of the borrowers (the $X_p$ dataset), and 16 variables describing the characteristics and evolution of the loan (the $X_f$ dataset). No applicant had a loan with the institution before, although some may have had loans with other institutions.

The dataset additionally included seven target variables: whether the customer defaulted in the first twelve months or not ($d_i$), whether the borrower defaulted at any time, and the time in case it did $(y_i, t_i)$, or whether there was no default during the history of the loan and how long was that history, and a pair of variables that indicates to what class ($C$ or $W$) the borrower had been assigned.

### 4.5.2 Economical Differentiation

The available variables that described the loan were:

- Amount: Amount granted to the borrower.
- Term: Term of the loan.
- Grace period (month): Grace period in months granted to the borrower.
- Grace period (day): Grace period in days granted to the borrower.

- Instalment Val.: Value of the instalments the borrower must pay.
- Collateral Val.: Value of the collateral the borrower had to give.
- Rate: Monthly rate the borrower was charged for the loan.
- Diff. In Rate: Difference between the maximum rate that loan could have been charged and the real rate charged. "Discount" applied.
- Prepay req.: If the borrower requested some prepayment to his/her loan.
- Prepaid: If the borrower effectively prepaid some of the amount of the loan.
- Renegotiated: If the loan comes from a renegotiation.
- Term Req.: Term the borrower requested originally for the loan.
- Instalment Req.: Term the borrower requested for the loan.
- Amount Paid: Total amount paid of the loan before default occurred.
- Arrears: Number of days the loan was in arrears before default occurred.
- Amount Req.: Amount requested by the borrower.

With these variables, the procedure described was run on the full database. The results from the clustering procedure can be seen in Table 4.1.

By contrast, Table 4.2 presents the results of a K-Means clustering of the same data. There are some interesting results: first, the K-Means clustering procedure shows important variations only in some variables, but these variations are more extreme. For example, the main split between the groups in Table 4.2 is in the variable Prepaid (whether the borrower prepaid some part of the loan or not), which is divided into those who did and those who did not, while in the constrained classification that variable is less important, but more interesting differences appear in the rest of the variables.

In particular, members of class $C$ receive better conditions for the loans, since the discount they received versus the maximum rate is higher, which may indicate a lower perceived risk by the lender. These members also try harder to repay the loan before defaulting, as can be seen in

Table 4.1: Results from Constrained Clustering of Defaulters.

| Variable | Class C | Class W |
|---|---|---|
| Amount | 530.238 | 338.118 |
| Term | 15,27 | 12,82 |
| Grace period (month) | 2,21 | 2,13 |
| Grace period (day) | 61,41 | 58,46 |
| Installment Val. | 43.829 | 32.656 |
| Collateral Val. | 1.576.543 | 435.231 |
| Rate | 2,25 | 2,14 |
| Dif. In Rate | -0,62 | -0,40 |
| Prepay req. | 0,85 | 0,18 |
| Prepaid | 0,44 | 0,07 |
| Renegotiated | 0,50 | 0,21 |
| Term Req. | 16,65 | 13,81 |
| Installment Req. | 55.788 | 43.707 |
| Amount Paid | 42.206 | 33.455 |
| Avg. Arrear | 111,41 | 73,14 |
| Amount Req. | 614.072 | 388.277 |

Table 4.2: Results from regular K-Means clustering of defaulters.

| Variable | Cluster 1 | Cluster 2 |
|---|---|---|
| Amount | 434.336 | 366.078 |
| Term | 14,70 | 12,80 |
| Grace period (month) | 2,22 | 2,11 |
| Grace period (day) | 61,52 | 57,97 |
| Installment Val. | 36.308 | 35.415 |
| Collateral Val. | 803.119 | 720.013 |
| Rate | 2,22 | 2,19 |
| Dif. In Rate | -0,63 | -0,36 |
| Prepay req. | 1,00 | 0,00 |
| Prepaid | 0,47 | 0,00 |
| Renegotiated | 0,53 | 0,15 |
| Term Req. | 15,86 | 13,86 |
| Installment Req. | 46.647 | 47.277 |
| Amount Paid | 34.438 | 36.709 |
| Avg. Arrear | 69,78 | 118,93 |
| Amount Req. | 491.797 | 426.709 |

variables Prepaid and Renegotiated, which indicates a better disposition to repay the loan, but an incapacity to do so since default occurred anyway. This may be attributed to a liquidity shock. There is a notable difference in the average number of days the borrower was in arrears before defaulting, more than 40 days, which again is a sign of more attempts to repay the loan. Finally, it can be seen that members in class $C$ request and receive longer terms than class $W$, receive higher amounts, and have to provide higher collaterals, but these variables where used in the constraints so some of the differences are due to these class definitions.

With these centroids, it is possible to split defaulters into two groups, creating a dataset with three classes that may be used to construct classification models. A borrower is assumed of belonging either to class $C$ or to class $W$, not both simultaneously, so the hypothesis is that a mixture model is the appropriate technique to be applied, since assuming pure competing risk would result in exaggerating the hazard the borrower is subject to at each time, as per equation (4.5). The next section shows the benchmarks and classification results.

### 4.5.3 Classification Results

Before applying the classification algorithm, it is necessary to select which variables are most relevant for each problem. A two step procedure was used, aimed at reducing the number of available variables and minimizing loss of useful information:

1. Univariate Filters: Each variable discriminating capabilities is calculated using the target variables $(d_i, y_i, y_i^C, y_i^W)$ in bivariate chi-squared or K-S tests, depending on whether the variables are nominal or continuous.

2. Over-adjusted trees: A series of classification trees were built on the dataset in order to determine which variables provided discrimination. The higher a variable appears in a tree, the more important its discriminative ability.

3. Wrapper forward and backward selection: Finally, for each model the variables where introduced one by one (forward procedure) or all were incorporated at once and removed one at a time (backward elimination). The procedure selects the more relevant variables for each target.

Table 4.3: Parameters for Competing Risks survival models.

(a) Class C

| Variables | *Beta* | P-value |
|---|---|---|
| Gender | -0.209 | 0.169 |
| Capital | 0.248 | 0.115 |
| Age | -0.991 | 0.010 |
| With Guarantor | -0.275 | 0.103 |
| Total Income | 2.066 | 0.016 |
| Previous Loans | -1.897 | 0.002 |
| Indirect Debt | 1.411 | 0.114 |

(b) Class W

| Variables | *Beta* | P-value |
|---|---|---|
| Gender | -0.211 | 0.017 |
| Married | 0.536 | 0.024 |
| Age | -1.064 | 0.000 |
| With Guarantor | -0.420 | 0.000 |
| Other Loans | -4.350 | 0.000 |
| Owner | -0.503 | 0.000 |
| Vehicle | -0.731 | 0.000 |
| Had Loan | -0.410 | 0.003 |
| Complies | -0.457 | 0.000 |
| Basic Services | -0.368 | 0.061 |
| Debt (Cons.) | 1.533 | 0.119 |

Finally, a set of 18 variables was used for constructing the models:

- Gender: Gender of the borrower.
- Marital Status: If the borrower is married, single or divorced.
- Age: Age in years of the borrower.
- With Guarantor: If the borrower has a guarantor for the debt.
- Other Loans: Number of loans the borrower reported to had had.
- Owner: If the borrower owns property.
- Vehicle: If the borrower owns a vehicle
- Had Loan: If the borrower had ever had a loan.
- Taxes: If the borrower pays taxes or is exempt (Declares) and if they are paid or not (Complies).
- Basic Services: If the basic services bills (water, electricity) had been paid on time the last three months.
- Debt (Cons.): Total debt in consumer products in the banking system.

Table 4.4: Parameters for Mixture Model.

(a) Survival Models

| Variable | Class *C* | Class *W* |
|---|---|---|
| Gender | 0.001 | -0.130 |
| Capitol | 0.473 | 0.203 |
| Age | 0.277 | -1.081 |
| Other Loans | -0.190 | -5.865 |
| Vehicle | 0.220 | -1.003 |
| Had Loan | -0.131 | -0.670 |
| Declares | 1.765 | 1.433 |
| Complies | -1.377 | -2.055 |
| Bills | 0.184 | -0.395 |
| Debt (Cons.) | 1.568 | 1.652 |

(b) Logistic Model

| Variable | Class *C* |
|---|---|
| Capitol | 0.055 |
| Age | -0.635 |
| With Guarantor | -0.083 |
| Total Sales | 0.873 |
| Other Loans | -0.907 |
| Declares | 0.459 |
| Complies | -0.360 |
| Total Debt | 0.266 |
| Indirect Debt | 0.961 |
| Indirect Leverage | 0.587 |
| Debt (Cons.) | 0.832 |

- Average Income: Average income of the borrower the last year.
- Capitol: If the borrower lives in the country's capitol.
- Bills: If all outstanding obligations of the borrower are up to date according to EQUIFAX database.
- Total Debt: Total debt of the borrower in the banking system.
- Indirect Debts: The total indirect debt of the borrower (Indirect) and the indirect debt deflated by the average income (Leverage). Indirect debt refers to debt for which the borrower is a guarantor.

In tables 4.3a and 4.3b the coefficients of the variables in the survival models for classes $C$ and $W$ can be observed, for the competing risk model, and in Table 4.4a and Table 4.4b the parameters for the survival models and the logistic part of the mixture model can be observed. In contrast, tables 4.5a and 4.5b show the coefficients of the survival and logistic regression models for which target only at default and non-default and do not split into the classes $C$ and $W$.

As expected, these latter two models have similar parameters, since the likelihood function are very similar, and the variables differ only in a slightly different target variable. Contrasting those results with the two regressions on the two default classes we get more interesting conclusions. First, the number of variables that are relevant in each dataset is different. Some variables (Gender,

Age, Guarantor, and Previous Loans) are relevant for all models, which indicates that are strong indicators of defaulting no matter the segmentation of defaulters, although interestingly Guarantor is only relevant in the logistic part of the mixture model, showing that the presence of guarantors is a signal that does not impact the time until an event occurs. Income is only relevant for class $C$ and the global model, but not for class $W$, which would make sense given the interpretation of class $W$: the liquidity of the borrowers is not the driver of default.

The variables describing the structure of the debt also bring interesting conclusions: In competing risks, it is the indirect debt (debt for which the borrower acts as a guarantor) which is more relevant for class $C$, while direct debt is more relevant for class $W$. One possible interpretation of this result is that for class $W$ whether the customer is currently repaying a loan in another institution is logically a driver of the willingness to repay. Regarding Class $C$, the mixture model shows a much more detailed profile in this respect: For the logistic function (Table 4.4b), all debts (direct and indirect) are relevant, which can be interpreted that borrowers in that class are more prone to default given *any* liquidity shock in the market, captured by the positive effects on all types of debt, but that defaulters in class $W$ are only more prone to default when their direct debt is considered.

The effect of Basic Services is also relevant only for the global model and class $W$. This is a strong indicator of willingness to repay, since Basic Services, whether the borrower has any debt with their basic services bills, represent a relatively small expense in the overall budget of the customer, so not paying these bills on time reflect obviously an effect in the level of financial proficiency of the borrower, which is actually what willingness to repay refers to. The same conclusion arises from Bills, the percentage of institutions in which the borrower has credit, but is up-to-date.

Regarding the differences between the Competing Risk model and the Mixture Model, the first noticeable one is the amount of variables that are significant for each model. In Competing Risks, Class $C$ model is only composed of seven variables, but the mixture model uses ten, and as will be seen below they both get to very similar global accuracies. competing risks, by using a single function for the class, is more sensitive to the fewer amount of defaulters that class $C$ has, so less variables are significant, whereas the Mixture Model uses a unique input matrix for both survival functions, so that sensitivity is diminished. It can be stated that the mixture model gives a more detailed account of the default event, at a cost of more predictive variables being used for the description, and that both the competing risk and the mixture model are more information-rich than the binary models, so there is a clear gain in understanding default from using a pure binary

Table 4.5: Parameters for Binary Models.

(a) Survival Model

| Variables | β | P-value |
|---|---|---|
| Gender | -0.207 | 0.007 |
| Married | -0.013 | 0.870 |
| Single | 0.486 | 0.019 |
| Age | -1.031 | 0.000 |
| With Guarantor | -0.370 | 0.000 |
| Previous Loans | -3.887 | 0.000 |
| Home Owner | -0.370 | 0.000 |
| Vehicle | -0.483 | 0.000 |
| Had Loans | -0.273 | 0.010 |
| Taxes | -0.390 | 0.000 |
| Basic Services | -0.393 | 0.019 |
| Average Income | 2.978 | 0.002 |

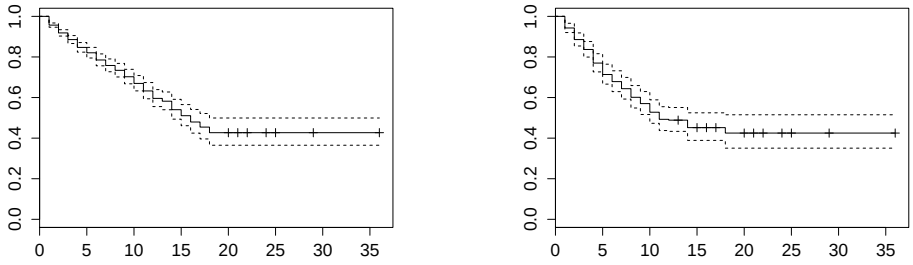
(b) Logistic Model

| Variable | β | P-value |
|---|---|---|
| Gender | -0.306 | 0.021 |
| Married | 0.084 | 0.553 |
| Single | 1.036 | 0.006 |
| Age | -0.028 | 0.000 |
| With Guarantor | -0.513 | 0.001 |
| Previous Loans | -0.950 | 0.000 |
| Home Owner | -0.402 | 0.006 |
| Vehicle | -0.692 | 0.000 |
| Had Loans | -0.445 | 0.005 |
| Complies | -0.005 | 0.001 |
| Basic Services | -0.377 | 0.005 |

approach, again at the cost of greater complexity.

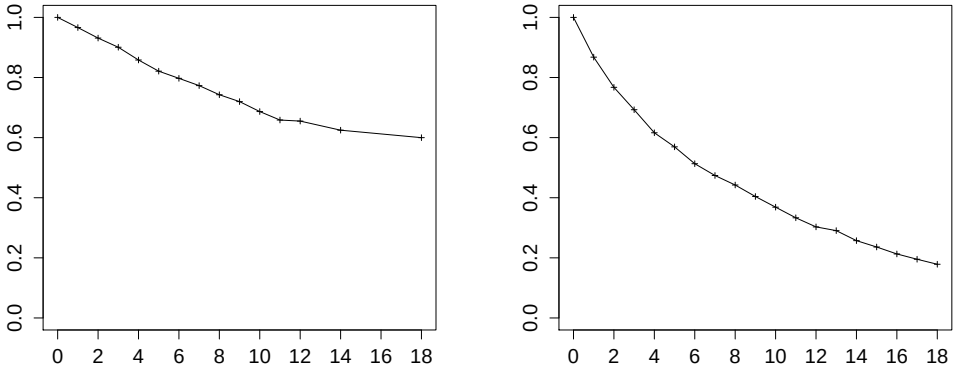
Another interesting result correspond to the survival curves. Since borrowers in class *C* are expected to default given exogenous, random circumstances, the survival curve should decrease constantly, without focusing in any particular instalment. On the other hand, borrowers in class *W* are expected to default much earlier, since the motivation to repay the loan would be lower. Such intuition is confirmed in figures 4.1a and 4.1b for Competing Risks, and Figures 4.2a and 4.2b for Mixture Models, which show precisely that: borrowers in class *C* present a straight decay, while borrowers in class *W* decay rapidly in the first five months, and in the next months this decay is softened. The difference between competing risks and mixture models is mainly that the former models an even survival curve for both cases (modelling absolute time to default for both classes), but the mixture construct a survival curve that is specific for each class. It can be observed in the latter that Class *W* has a much higher default rate in the long term.

Finally, the classification improvements in using the two default classes can be seen in table 4.6. Although minor, there is an observable improvement between using the three classes models against using only two. The small difference may be influenced by the small number of defaulters in the database, since with only 300 cases in class *C* it is more complicated to obtain greater differences. Another argument in favour of using competing risks and mixture models is that there is a temporal difference in this discriminating capacity, considering that for the first five months there is a much marked difference in the hazard rate between borrowers of class *C* and class *W*, which could bring




(a) Survival for Class $C$ (b) Survival for Class $W$

Figure 4.1: Survival curves for competing risks model. There is a much more rapid decay in class $W$, as expected.




(a) Survival for Class $C$ (b) Survival for Class $W$

Figure 4.2: Survival curves for mixture model. Again class $W$ rapidly decays, but now the models detect the greater default rate they have.

improvements in provisioning, for example. However, as was mentioned above, the higher cost that arises from the need to collect a larger amount of information for the multi-class models can be a detriment for the implementation of such models.

When comparing the accuracy of both types of multi-class models, actually competing risks is the slightly superior model. The additive hazard assumption does not seem to impact on the discriminative capacity of the model, which could mean that the assumption still allows for the model to reach the Cramer-Rao lower bound for predictive capability. The added complexity of the Mixture Model then is useful for a better understanding of the event, but that does not imply a

Table 4.6: Area Under the Curve (AUC) of the Tested Models

| Model | AUC |
|---|---|
| Binary Survival | 0.7600 |
| Competing Risks | 0.7663 |
| Logistic Regresion | 0.7595 |
| Mixture Model | 0.7612 |

better discriminative capacity in this case.

## 4.6 Conclusions

In this paper we study the use of multi-class models in credit scoring, focusing on two types: Competing Risks and Mixture Models. Both multi-class models differ on the basic assumption regarding the hazard rates: Competing Risks identifies the hazard rate as an additive function among classes, resulting a survival function that resembles fully independent events. Mixture Models, in contrast, build a survival function from a time-independent distribution for each class, and a time dependent model that is also unique for each type of default, resulting in a more complex relationship between the events.

Multiple classes of defaulters can arise from different operational reasons, but in this paper we construct them from a procedure that consisted on a two step procedure which first performs economic modelling of defaulters, and then incorporates this modelling as domain-knowledge on a semi-supervised clustering methodology. The procedure was tested on previously unpublished data from approximately 5,000 low-to-middle income micro-entrepreneurs from a Chilean private institution.

Survival analysis models are relevant in the context of credit scoring since they provide a wider approach to the problem of credit risk, incorporating a temporal variable into the analysis. This allows for more flexibility in the decision making process when using them. In this work we present a general framework for the more general case when there are more than two classes of

borrowers, showing it is possible to extend the classification methodologies using what is available in medical literature.

The results are useful since they create a profile of borrowers regarding their expected behaviour when observing default: separating borrowers due to problems in capacity of repayment, and problems in willingness to repay. Experimental results in a real database show the numeric evidence that is extracted from this methodology. Customers that are flagged as with problems in capacity try harder to repay the loan, have more days in arrears, prepay more, or ask for benefits such as renegotiation. The lender also seems to have identified them at origination, since the borrowers got better conditions than the other class. These results hint at the existence of two classes of defaulters, and that they can be profiled using existing techniques from data mining and economic finance.

The methodology can also be used for classification, and the results using multi-class survival analysis showed that customers in the class flagged as not wanting to repay fail earlier than the other borrowers, with more rapidly decaying survival curves. Additionally, three sets of variables were found relevant in the models: a set that is common to both classes, and two sets that model willingness to repay and capacity to repay, which were relevant only to one segment. The results, when measured in AUC, present a slight advantage for the multi-class models. Identifying the profile of defaulters via the clustering method can bring additional gains when defining the policy for accepting customers, which may improve classification further.

When contrasting the different approaches for multi-class survival analysis, Mixture Models arrive at a much more detailed characterization of the default event, but at a cost of higher number of predictive variables, and there does not appear to be a noticeable gain in predictive capability. Furthermore, we found in this case that the additive hazard rate that Competing Risks uses, while theoretically imprecise, does not negatively impact the obtained procedure in predictive capability. The survival curves that result from each model are also different, with the Mixture Model representing a more class-dependant curve, and the Competing Risks model representing an overall curve that is different among classes only in their convexity, being more pronounced for Class $W$. The decision of which model to use is then related to a trade-off between operational costs and defaulter description: Competing Risks results is comparatively inexpensive, but the Mixture Model brings more detailed profile descriptions of defaulters.

Future work in this area would include further understanding defaulter behaviour to arrive at

better economic models of the reasons for default. These models could be incorporated in multi-class and binary survival analysis methods to improve provisioning and classification.

## 4.7 Acknowledgment

The first author acknowledges the Chilean National Council for Research, Science and Technology (CONICYT) for the grant that supports this work (AT-24110006), the PhD in Engineering Systems for its support, and to S. Beckman and C. Mora for their editing work. This paper has been partially funded by the Institute of Complex Engineering Systems (ICM: P-05-004-F, CONICYT: FBO16) and the Finance Center at the Department of Industrial Engineering, Universidad de Chile.

# Bibliography

ALARY, D., GOLLIER, C. Debt contract, strategic default, and optimal penalties with judgement errors. Annals of Economics and Finance, 5:357–372, 2004.

ALLEN, L., ROSE, L. C. Financial survival analysis of defaulted debtors. Journal of the Operational Research Society, 57(6):630–636, 2006.

ANDERSON, R. The Credit Scoring Toolkit. Oxford University Press, 2007.

ARROW, K. J., DEBREU, G. Landmark Papers in General Equilibrium Theory, Social Choice and Welfare. Edward Elgar Publishing, 2002.

BAESENS, B., VAN GESTEL, T., STEPANOVA, M., VANTHIENEN, J. Neural network survival analysis for personal loan data. Journal of the Operational Research Society, 56(9):1089–1098, 2005.

BAESENS, B., VAN GESTEL, T., VIAENE, S., STEPANOVA, M., SUYKENS, J., VANTHIENEN, J. Benchmarking state-of-the-art classification algorithms for credit scoring. The Journal of the Operational Research Society, 54(6):627–636, 2003.

BANASIK, J., CROOK, J. C., THOMAS, L. C. Not if but when will borrowers default. Journal of the Operational Research Society, 50(12):1185–1190, 1999.

BARD, J. F., JARRAH, A. . Large-scale constrained clustering for rationalizing pickup and delivery operations. Transportation Research Part B: Methodological, 43(5):542–561, 2009. URL http://econpapers.repec.org/RePEc:eee:transb:v:43:y:2009:i:5:p:542-561.

BASU, S., BILENKO, M., MOONEY, R. J. A probabilistic framework for semi-supervised clustering. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '04, pp. 59–68. New York, NY, USA: ACM, 2004. URL http://doi.acm.org/10.1145/1014052.1014062.

BASU, S., DAVIDSON, I., WAGSTAFF, K. Constrained Clustering: Advances in Algorithms, Theory, and Applications. Chapman & Hall/CRC, 2008.

BELLOTTI, T., CROOK, J. Credit scoring with macroeconomic variables using survival analysis. Journal of the Operational Research Society, 60(12):1699–1707, 2008.

BENTON, M., MEIER, S., SPRENGER, C. Overborrowing and undersaving: lessons and policy implications from research in behavioral economics. Public and Community Affairs Discussion Papers 2007-4, Federal Reserve Bank of Boston, 2007.

BENZION, U., RAPOPORT, A., YAGIL, J. Discount rates inferred from decisions: An experimental study. Management Science, 35(3):pp. 270–284, 1989. URL http://www.jstor.org/stable/2631972.

BLOCHLINGER, A., LEIPPOLD, M. Economic benefit of powerful credit scoring. Journal of Banking & Finance, 30(3):851–873, 2006.

BLOCK-LIEB, S., JANGER, E. J. The myth of the rational borrower: Rationality, behaviorism, and the misguided "reform" of bankruptcy law. Texas Law Review, 84:1481–1565, 2006.

BRAVO, C., FIGUEROA, N., WEBER, R. Game theory and data mining model for price dynamics in financial institutions. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN 2010), pp. 1 – 8. 2010.

BRAVO, C., THOMAS, L. C., WEBER, R. Improving credit scoring by differentiating defaulter behavior. Journal of the Operational Research Society, -:Accepted for Publication, 2013.

BRAVO, C., WEBER, R. Semi-supervised constrained clustering with cluster outlier filtering. In: MARTIN, C. S., KIM, S.-W. (Eds.) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Lecture Notes in Computer Science 7042, pp. 347–354. Springer Verlag, 2011.

BRESLOW, N. E. Covariance analysis of censored survival data. Biometrics, 30:89–100, 1974.

BROWN, I., MUES, C. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, 39(3):3446–3453, 2012. URL http://www.sciencedirect.com/science/article/pii/S095741741101342X.

BURKS, S. V., CARPENTER, J. P., GOETTE, L., RUSTICHINI, A. Cognitive skills explain economic preferences, strategic behavior, and job attachment. IZA Discussion Papers 3609, Institute for the Study of Labor (IZA), 2008. URL http://ideas.repec.org/p/iza/izadps/dp3609.html.

CASTERMANS, G., MARTENS, D., VAN GESTEL, T., HAMERS, B., BAESENS, B. An overview and framework for pd backtesting and benchmarking. Journal of the Operational Research Society, 61:359–373, 2010.

CHABRIS, C. F., LAIBSON, D., MORRIS, C. L., SCHULDT, J. P., TAUBINSKY, D. Individual laboratory-measured discount rates predict field behavior. NBER Working Paper No. 14270, 2008. URL http://www.nber.org/papers/w14270.

CIESLAK, D., CHAWLA, N. Detecting fractures in classifier performance. In: Proceedings of the Seventh IEEE International Conference on Data Mining, pp. 123–132. Department of Computer Science and Engineering, University of Notredame, 2007.

COX, D. R. Regression models and life tables. Journal of the Royal Statistical Society. Series B (Methodological), 34(2):187–220, 1972.

CROWDER, M. J. Classical Competing Risks. Chapman and Hall/CRC, 2001, 200p pp.

DAVIDSON, I., RAVI, S. S. Clustering with constraints: Feasibility issues and the k-means algorithm. In: Proceedings of the SIAM International Conference on Data Mining (SDM 2005). 2005.

DE WAAL, D. A., DU TOIT, J. V., DE LA REY, T. An investigation into the use of generalized additive neural networks in credit scoring. In: Proceedings of IX Credit Scoring & Credit Control Conference. Pollock Halls, University of Edinburgh, Scotland, 2005. URL http://www.crc.man.ed.ac.uk/conference/archive/2005/papers/de-waal-du-toit-de-la-rey.pdf.

DIVISIÓN EMPRESAS DE MENOR TAMAÑO, M. d. E. Encuesta longitudinal de empresas [longitudinal survey of companies]. Retrieved online 14 February, 2012., 2009. URL http://www.observatorioempresas.gob.cl/LinkClick.aspx?fileticket=ACOoveG2z70%3d&tabid=63.

DOĞAN, H., GÜZELIŞ, C. Gradient networks for clustering. In: GÖKNAR, I. C., SEVGI, L. (Eds.) Complex Computing-Networks, vol. 104 of *Springer Proceedings Physics*, pp. 275–278. Springer Berlin Heidelberg, 2006. URL http://dx.doi.org/10.1007/3-540-30636-6_30. 10.1007/3-540-30636-6.

EFRON, B. The efficiency of cox's likelihood function for censored data. Journal of American Statistical Association, 72:557–565, 1977.

ESCARELA, G., BOWATER, R. J. Fitting a semi-parametric mixture model for competing risks in survival data. Communications in Statistics: Theory and Methods, 37(2):277–293, 2008.

FAYYAD, U., PIATETSKY-SHAPIRO, G., SMYTH, P. The KDD process for extracting useful knowledge from volumes of data., , . Communications of the ACM, 39(11):27–34, 1996.

FINLAY, S. Multiple classifier architectures and their application to credit risk assessment. European Journal of Operational Research, 210:368–378, 2011.

FUDENBERG, D., TIROLE, J. Game Theory. MIT Press, 1991.

GLENNON, D. C., NIGRO, P. Measuring the default risk of small business loans: A survival analysis approach. Journal of Money, Credit, and Banking, 37(5):923–947, 2005.

GREEN, L., FRY, A. F., MYERSON, J. Discounting of delayed rewards: A life-span comparison. Psychological Science, 5(1):pp. 33–36, 1994. URL http://www.jstor.org/stable/40062338.

GREENE, W. H. Econometric Analysis. Prentice Hall, 1993.

GUISO, L., SAPIENZA, P., ZINGALES, L. The determinants of attitudes towards strategic default on mortgages. Economics Working Papers ECO2010/31, European University Institute, 2010. URL http://econpapers.repec.org/RePEc:eui:euiwps:eco2010/31.

HAND, D., HENLEY, W. Statistical classification methods in consumer credit scoring: a review. Journal of the Royal Statistical Society Association, 160:523–541, 1997.

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd edition ed. Springer, 2009, 552p pp.

HOSMER, D., LEMESHOW, H. Applied Logistic Regression. John Wiley & Sons, 2000.

HOSMER, D., LEMESHOW, S., MAY, S. Applied Survival Analysis. John Wiley & Sons, 2008.

INSTITUTO NACIONAL DE ESTADÍSTICAS, I. Primera encuesta de las micro, pequeñas y medianas empresas [first survey of micro, small, and medium companies]. Retrieved online 14 February, 2012., 2002. URL http://www.ine.cl/canales/chile_estadistico/estadisticas_economicas/pymes/pdf/resultadospyme.pdf.

JOKIVUOLLE, E., PEURA, S. A model for estimating recovery rates and collateral haircuts for bank loans. Research Discussion Papers 2/2000, Bank of Finland, 2000. URL http://ideas.repec.org/p/hhs/bofrdp/2000_002.html.

KALBFLEISCH, J. D., PRENTICE, R. L. The Statistical Analysis of Failure Time Data. 2nd edition ed. John Wiley & Sons, 2002.

KAPLAN, E. L., MEIER, P. Non-parametric estimation with incomplete observations. Journal of the American Statistical Association, 53:457–481, 1958.

KIM, H. S., SOHN, S. Y. Random effects logistic regression model for default prediction of technology credit garantee fund. European Journal of Operational Research, 183:472–478, 2007.

KIM, H. S., SOHN, S. Y. Support vector machines for default prediction of SMEs based on technology credit. European Journal of Operational Research, 201:838–846, 2010.

KUK, A. Y. C., CHEN, C.-H. A mixture model combining logistic regression with proportional hazards regression. Biometrika, 79:531–541, 1992.

LARSON, M. G., DINSE, G. E. A mixture model for the regression analysis of competing risks data. Journal of the Royal Statistical Society. Series C (Applied Statistics), 34(3):201–211, 1985.

LEVY, M., SANDLER, M. Structural segmentation of musical audio by constrained clustering. Audio, Speech, and Language Processing, IEEE Transactions on, 16(2):318 –326, 2008.

L'HUILLIER, G., WEBER, R., FIGUEROA, N. Online phishing classification using adversarial data mining and signaling games. SIGKDD Explor. Newsl., 11:92–99, 2010. URL http://doi.acm.org/10.1145/1809400.1809421.

LIN, S., ANSELL, J., ANDREEVA, G. Predicting default of a small business using different definitions of financial distress. The Journal of the Operational Research Society, Advance Online Publication:10 August, 2011.

MACKINON, M. J., GLICK, N. Data mining and knowledge discovery in databases - an overview. Australian & New Zealand Journal of Statistics, 41(3):255–275, 1999.

MALIK, M., THOMAS, L. C. Modelling credit risk of portfolio of consumer loans. Journal of the Operational Research Society, 61(3):411–420, 2009.

MARUBINI, E., VALSECCHI, M. G. Analysing Survival Data from Clinical Trials and Observational Studies. John Wiley & Sons, 2004.

MINISTERIO DE ECONOMÍA. Segunda encuesta de microemprendimiento [second micro-entrepreneurship survey]. Retrieved online 5 May, 2012, 2012. URL http://www.economia.gob.cl/wp-content/uploads/2012/04/Informe-de-resultados-seleccionados-EME2.pdf.

MOFFAT, P. G. Hurdle models of loan default. The Journal of the Operational Research Society, 56(9):1063–1071, 2005.

MYATT, G. J., JOHNSON, W. P. Making Sense of Data II. John Wiley & Sons, 2009.

NARAIN, B., THOMAS, L. C., CROOK, J. N., EDELMAN, D. B. Credit scoring and credit control, chap. Survival analysis and the credit granting decision, pp. 109–121. Oxford, UK, 1992.

NG, S. K., MCLACHLAN, G. J. An em-based semi-parametric mixture model approach to the regression analysis of competing-risks data. Statistics in Medicine, 22(7):1097–1111, 2003.

OZDEMIR, B., MIU, P. Basel II Implementation. McGraw-Hill, 2009.

PATIL, G., MODARRES, R., MYERS, W., PATANKAR, P. Spatially constrained clustering and upper level set scan hotspot detection in surveillance geoinformatics. Environmental and Ecological Statistics, 13:365–377, 2006. URL http://dx.doi.org/10.1007/s10651-006-0017-5. 10.1007/s10651-006-0017-5.

PENG, Y. Fitting semiparametric cure models. Computational Statistics & Data Analysis, 41(3-4):481–490, 2003.

SAFAVIAN, S. R., LANDGREVE, D. A survey of decision tree classifier methodology. IEEE Transactions on Systems, Man, and Cybernetics, 21(3):660–674, 1991.

SCHREINER, M. Credit scoring for microfinance - can it work? Journal of Microfinance, 2(2):105–119, 2000.

SETIONO, R., BAESENS, B., MUES, C. A note on knowledge discovery using neural networks and its application to credit card screening. European Journal of Operational Research, 192(1):326–332, 2009.

SIDDIQI, N. Credit risk scorecards: developing and implementing intelligent credit scoring. John Wiley and Sons, 2006, 196p pp.

STEPANOVA, M., THOMAS, L. C. Survival analysis methods for personal loan data. Operations Research, 50(2):277–289, 2002.

STIGLITZ, J. E., WEISS, A. Credit rationing in markets with imperfect information. American Economic Review, 71(3):393–410, 1981. URL http://ideas.repec.org/a/aea/aecrev/v71y1981i3p393-410.html.

SUPERINTENDENCIA DE BANCOS E INSTITUCIONES FINANCIERAS. Compendio de Normas Contables [Compendium of Accounting Rules]. SBIF, 2008, 169p pp. URL http://www.sbif.cl/sbifweb/servlet/Biblioteca?indice=C.D.A\&idContenido=6742.

THOMAS, L. C. A survey of credit and behavioral scoring: forecasting financial risk of lending to consumers. International Journal of Forecasting, 16(2):149–172, 2000.

THOMAS, L. C., CROOK, J. N., EDELMAN, D. B. Credit Scoring and its Applications. SIAM, 2002.

TONG, E. N. C., MUES, C., THOMAS, L. C. Mixture cure models: if and when borrowers default. European Journal of Operations Research, 218(1):132–139, 2012.

VAN DER BURGT, M. Calibrating low-default portfolios, using the cumulative accuracy profile. Journal of Risk Model Validation, 1(4):17–33, 2008.

VAN GOOL, J., VERBEKE, W., SERCU, P., BAESENS, B. Credit scoring for microfinance: is it worth it? International Journal of Finance & Economics, Available Online:Accepted for Publication, 2011.

VATSA, V., SURAL, S., MAJUMDAR, A. A game-theoretic approach to credit card fraud detection. In: JAJODIA, S., MAZUMDAR, C. (Eds.) Information Systems Security, vol. 3803 of *Lecture Notes in Computer Science*, pp. 263–276. Springer Berlin / Heidelberg, 2005. URL http://dx.doi.org/10.1007/11593980_20.

WAGSTAFF, K., CARDIE, C., ROGERS, S., SCHROEDL, S. Constrained k-means clustering with background knowledge. In: In Proceedings of the Eighteenth International Conference on Machine Learning, pp. 577–584. Morgan Kaufmann, 2001.

WANG, Y. Combining data mining and game theory in manufacturing strategy analysis. Journal of Intelligent Manufacturing, 18:505–511, 2007. URL http://dx.doi.org/10.1007/s10845-007-0054-4. 10.1007/s10845-007-0054-4.

WETTE, H. C. Collateral in credit rationing in markets with imperfect information: Note. The American Economic Review, 73(3):442–445, 1983.

WHITE, L., SMITH, H., CURRIE, C. OR in developing countries: A review. European Journal of Operational Research, 208:1–11, 2011.

XU, R., WUNSCH, D. Clustering. Wiley-IEEE Press, 2008.

YAMASHITA, S., YOSHIBA, T. Analytical solution for expected loss of a collateralized loan: A square-root intensity process negatively correlated with collateral value. Discussion Paper 2010-E-10, Bank of Japan, 2010.

ZEIRA, G., LAST, M., MAIMON, O. Advanced Techniques in Knowledge Discovery and Data Mining, chap. Segmentation on Continuous Data Streams Based on a Change Detection Methodology, pp. 103–126. Springer, 2005.

ZHANG, G. P. Avoiding pitfalls in neural network research. IEEE Transactions on Systems, Man, and Cybernetics, Part C, 37(1):3–16, 2007.

ZHAO, W., HE, Q., MA, H., SHI, Z. Effective semi-supervised document clustering via active learning with instance-level constraints. Knowledge and Information Systems, pp. 1–19, 2011. URL http://dx.doi.org/10.1007/s10115-011-0389-1. 10.1007/s10115-011-0389-1.

ZHU, H., BELING, P. A., OVERSTREET, G. A. A study in the combination of two consumer credit scores. The Journal of the Operational Research Society, 52(9):974–980, 2001.

# Appendix A

# Granting and Managing Loans for Micro-Entrepreneurs: New Developments and Practical Experiences[1]

**Cristián Bravo, Sebastián Maldonado, and Richard Weber**

**Abstract**

We present a methodology to grant and follow-up credits for micro-entrepreneurs. This segment of grantees is very relevant for many economies, especially in developing countries, but shows a behaviour different to that of classical consumers where established credit scoring systems exist. Parts of our methodology follow a proven procedure we have applied successfully in several credit scoring projects. Other parts, such as cut-off point construction and model follow-up, however, had to be developed and constitute original contributions of the present paper. The results from two credit scoring projects we developed in Chile, one for a private bank and one for a governmental credit granting institution, provide interesting insights into micro-entrepreneurs' repayment behaviour which could also be interesting for the respective segment in countries with similar characteristics.

[1]The following is a pre-print version of the work submitted to European Journal in Operational Research, currently in second review.

# A.1 Introduction

Credit scoring corresponds to the use of statistical models to transform relevant data into numerical measures that guide credit decisions (Anderson, 2007), and its main objective is to estimate the probability of default, i.e. the event of a customer not paying back the loan in a given time period. Recent developments in credit scoring are oriented, for example, in analysing imbalanced credit datasets (Brown and Mues, 2012), in survival analysis (Bellotti and Crook, 2008; Tong et al., 2012), in correct ways to validate credit scoring models and make them comprehensible (Castermans et al., 2010), and in adapting new models for credit scoring use (Setiono et al., 2009).

This work focuses on micro-entrepreneurs, a segment different from independent persons or large companies in terms of size, income, and organizational structure. Although several efforts have been undertaken to gain knowledge about the default risk of small and medium-sized enterprises – see, e.g., Kim and Sohn (2010), or Van Gool et al. (2011) –, only few studies are available for micro-entrepreneurs (Schreiner, 2000).

The aim of this paper is two-fold: First, it provides experiences and insights we gained from several credit scoring projects – a private bank and a state-owned organization – where we had to adapt existing methodologies for credit granting. Secondly, these projects required the development of new techniques for credit scoring, namely cut-off point construction and model follow-up, which will be described subsequently.

This paper is organized as follows. The next section characterizes Chilean micro-entrepreneurs. Subsequently, we present the methodology used to construct credit scoring models putting special emphasis on stages where problems arose or special knowledge was revealed. The following section contains the results we obtained applying the proposed methodology to the state-owned organization. Finally, conclusions are drawn from our work in Section A.6.

# A.2 Chilean Micro-entrepreneurs

In most countries, micro-entrepreneurs are an important element of the economy, accounting for the creation of new business opportunities and employment. Chilean micro-entrepreneurs are defined as very small firms, with up to ten employees and an annual income of no more than EUR 97,000. They offer 21% of the jobs in the country (División Empresas de Menor Tamaño, 2009), and represent a large portion of Chilean companies), but they generate only a small portion of annual sales, with micro-entrepreneurs representing

13.9% of total annual sales of all companies in the country (Instituto Nacional de Estadísticas, 2002).

Micro-entrepreneurs, usually receiving support only from governments and not-for-profit initiatives, have become attractive customers for banks and loan-granting institutions in recent years, but the risk measures that accompany them have to follow suit. In particular, micro-entrepreneurs have certain characteristics that must be taken into account when developing a credit scoring system. For example, they are usually on a tight budget, regardless of their revenues. So it is not a question of how much money they make, which takes away a natural candidate for a discriminatory variable in credit scoring models. Considering the special characteristics that micro-entrepreneurs possess, there is still little knowledge on the variables that may determine their risk as loan borrowers. Additionally, most micro-entrepreneurs have not had any access to financial instruments previously, so the common scorecards that are available from major distributors are concentrated on higher risk segments (due to lack of information), and makes the segment be considered as "high risk" without any deeper consideration (Ministerio de Economía, 2012).

The question may arise of whether funding is even needed for this segment. A recent government study (Ministerio de Economía, 2012) found that 81 percent of micro-entrepreneurs fund their start-ups using their personal savings or family loans. Only four percent of the start-ups were funded using banks or established financial institutions, reporting that the access to credit is given only by supermarkets and some retailers (in the form of personal credit cards). The main conclusion we draw from the information presented is that loans are necessary for this segment, but the perceived risk of the micro-entrepreneur is too high for traditional financial institutions to provide coverage. We believe that existing credit scoring methodologies must be adapted in order to correctly reflect the reality of micro-entrepreneurs and, in that way, create the conditions for an equal-opportunity and profitable environment both for micro-entrepreneurs and for the institutions that grant them loans.

## A.3 Developing a Credit Scoring System

We applied the KDD process (Fayyad et al., 1996) to develop both models; one for the private bank and the other for the state-owned organization. In the following subsection we describe the process of data acquisition and consolidation, followed by the process of data cleansing. Finally we present the process for estimating the probability of default for a solicited loan.

### A.3.1 Data Set Consolidation

The first steps were to identify the relevant databases in which the information was scattered, extract the respective variables, and load them into a repository especially created for the construction of our models.

In order to develop an effective credit scoring system, a homogeneous and representative sample of the population for both classes (defaulters and non-defaulter) is needed. Based on the relevant literature, e.g. Thomas et al. (2002), and our experience from other similar credit scoring projects, we first segmented customers and the requested loans by differentiating between new customers in one group and renewing or current customers in the other.

### A.3.2 Data Cleansing and Variable Selection

The next step consisted of data preparation and variable (feature) selection. The procedure applied had two goals: to select features in a cascade-like approach thus minimizing the risk of eliminating potentially useful ones, and to maximize the knowledge extracted from the respective dataset. The description of the procedure follows:

1. Concentration of feature values and analysis of missing values: In order to quickly discard useless variables, those concentrated in a single value in more than 99% of the cases, and those with more than 30% of missing values were eliminated. The rationale of the second criterion is to reduce the number of discarded cases or imputations realized in order to construct the final model.
2. Univariate analysis: The remaining variables were tested to find distribution equality across groups, using the objective variables (defaulter / non-defaulter) as the splitting criterion. In particular Kolmogorov-Smirnov (K-S) and $\chi^2$-tests were applied to continuous and discrete variables, respectively.
3. Final data preparation: The selected variables accounted for 20% of the original ones. The resulting dataset had very few null cases (less than 1%), which were eliminated, and resulted in a robust subset of the variables which were then used as input for a logistic regression model.

### A.3.3 Estimating the Probability of Default

The final decision regarding the acceptance of the loan application can be made by comparing the calculated probability of default, estimated using a suitable model and the variables obtained from previous steps, with a suitable pre-defined threshold (Hand and Henley, 1997). Several studies in credit scoring have focused on comparing the classification performance of different techniques (see, for example, Baesens et al. (2003), Finlay (2011)). Their main conclusion is that the traditional credit scoring method logistic regression reaches performance levels comparable with more sophisticated data mining approaches for modelling credit risk, being therefore the main technique used for credit scorecard construction (Thomas et al., 2002).

The objective of logistic regression is to construct a function which determines the probability of default for a given client. It considers $V$ different regressors in a vector $\mathbf{x}_i \in \mathbb{R}^V, i = 1,...,N$, and an observed binary variable $y_i = 1$ if borrower $i$ defaults, and $y_i = 0$ else. Considering the dependent variable as latent, the probability of default is:

$$p(y_i = 1|\mathbf{x}_i) = \frac{1}{1+e^{-\left(\beta_0+\sum_{j=1}^{V}\beta_j x_{ij}\right)}}, \tag{A.1}$$

where $\beta_0$ is the intercept, and $\beta_j$ is the regression coefficient associated to variable $j$. Since these parameters are unknown, they have to be estimated using, e.g., a maximum likelihood algorithm, which results in unbiased, asymptotically normally distributed estimators $\hat{\beta}_j$ (Hosmer and Lemeshow, 2000). The expression in the exponent is a measure of the total contribution of the independent variables used in the model, and is known as the logit (Greene, 1993).

The process of constructing the model uses a wrapper for variable selection, consisting of a greedy search of variables that accounts for maximum discriminatory capacity. There are two common approaches (Hosmer and Lemeshow, 2000): *forward selection* and a *backward elimination*. Both methods are based on measuring how relevant each variable is, as measured by a $\chi^2$-test.

## A.4 New Developments

The final step when constructing a credit scoring system is deciding the cut-off to be used in order to transform the probability obtained from the logistic regression into a binary decision. However, when profits are not well-defined, this is not a trivial task as we found, e.g., in our project with the state-owned, non-profit institution. We developed a generic methodology to determine such cut-off points for the case of non-profit organizations. Additionally, once the credit scoring model is being used, different kinds of changes might occur diminishing predictive capability. We developed procedures to follow such shifts in the population and update the respective model parameters accordingly. These approaches are presented in the subsequent subsection.

### A.4.1 Methodology for Cut-Off Point Construction

The methodology to determine a cut-off point starts by determining the cost of accepting a bad applicant using the expected loss for a given loan. The second part is estimating the cost of rejecting a good applicant, point especially interesting for any governmental institution, usually more interested in public wealth than in monetary profits. Finally, both numbers are considered and combined to calculate an optimal cut-off point to be applied to the result of the logistic regression model.

#### A.4.1.1 Cost of accepting a bad applicant

Our cut-off point methodology starts by segmenting the range of the possible values the estimated default probability can take. Considering the database of loans that were used to validate the model, it is necessary to obtain the probabilities of default for each customer. Then these values are ordered and segmented in ranges depending on the size of the database. In our experience, intervals of 0.05 turned out to provide the best results, generating 20 segments of the interval [0.1].

The direct cost for the organization when a grantee does not repay his or her loan has to be calculated by taking into consideration all the resources owed. Additionally, the value of the collateral must be discounted, that is, the real estate or different properties that the grantee declares as security for the loan. In general, the loss per loan (given that default occurred) is the product of the following quantities(Ozdemir and Miu, 2009):

- EAD: Exposure at default. Is the amount that the grantor is owed when default occurs, including all standing instalments and any owed interest. In the case of loans with guarantors, the value of the loss and the exposure is different (Superintendencia de Bancos e Instituciones Financieras, 2008), but for this particular case (no guarantors considered) they are assumed to be the same.
- LGD: Loss Given Default. Proportion of the exposed capital (EAD) that is actually lost given the event of default. This value considers the expected proportion of the loan that will not be paid by the customer after default occurs including the recovery after prosecution or collection, and the recovery given by the collateral.

The cost of defaulting for each customer follows $LOSS = LGD \cdot EAD$. In order to determine the cut-off point it is necessary to know the accumulated cost for all customers. The final amount corresponds to the set of defaulters whose default probabilities are below the cut-off point $p$:

$$C^{p}_{\text{Loss}} = \frac{\sum_{i \in D(p)} LOSS_i}{|D(p)|} \qquad p \in \{0.05, 0.1, ..., 0.95, 1\}, \tag{A.2}$$

where $D(p)$ is the set of defaulters with estimated probability of default less than $p$, $|D(p)|$ is the number of customers that belong to set $D(p)$, and $LOSS_i$ is the observed loss for customer $i$.

The cost is divided by $|D(p)|$ so an average cost per loan in that cut-off is determined. This way, when applying the model, corrections can be introduced considering the actual number of loans that are observed – a new $|D(p)|$ – tying the cut-off to market conditions.

#### A.4.1.2 Cost of Rejecting a Good Applicant

The case of a private loan-granting company rejecting a potentially good customer leads to an opportunity cost equivalent to the gain or utility the loan would have generated for the financial entity. This cost has an associated market share loss: if the financial policy has been too restrictive and many potentially good borrowers have been rejected, the company exposes itself to a commercial risk by reaching fewer customers than would have been possible.

In the case of a state-owned institution, however, the cost of rejected loans has to consider the profit a rejected loan could generate for the organization (opportunity cost) plus the benefit for society that is not going to occur since the credit is not granted. In general, this benefit is not relevant for private organizations

Table A.1: Comparison of Annual Income for Customers and Control Group.

| | With Loan | Without Loan | Diff./Avg. | Percentage |
|---|---|---|---|---|
| Avg. annual income | 4,869 EUR | 3,582 EUR | 1,287 EUR | 26.44% |
| Est. loss of income (if loan not granted) | 1,247 EUR | | 1,314 EUR | 26.97% |
| Est. additional income (if loan granted) | | 1,377 EUR | | |

when determining their cut-off points. Furthermore, it could be argued that the profits that any institution (public or private) perceives is at the expense of the borrower, so the net social benefit of profits is zero, therefore should not be considered.

In order to estimate this lost benefit for society, we compared the average income of applicants who **did** receive loans from the institution with average income of entrepreneurs who **did not** receive any loan. A previous study provided by the government institution analysed a sample of 1,010 entrepreneurs who received a loan and a sample of 500 workers who did not receive any loan (control group).

The survey was conducted in the year 2005, considering granted loans between 2000 and 2003 for the study group. These granted loans were repaid before the survey was taken. The survey was controlled by region and activity by using cluster sampling, but also another variables were studied in order to validate the sampling process, such as age, sex, size of the family group, ethnic group, and educational level. No significant differences were found for these variables.

The average annual income obtained by the customers (study group) was 4,869 EUR, which is 1,287 EUR higher than the average of the control group (3,582 EUR). Additionally, the customers who received a loan were asked to estimate their expected loss in terms of annual income in case they would not have received credit. Those who did not receive a loan were asked to estimate how much additional income they would have perceived. The averages of both estimations are similar to the difference of the average annual income for both groups (1,314 EUR and 1,287 EUR respectively), making the estimation robust. Table A.1 summarizes the results of this analysis.

According to this information, the average income of applicants with access to a loan is 26.44% higher than the ones without this form of financial help. Given that the vast majority of these entrepreneurs without access to loans do not get credit from private financial institutions and tend to reduce their production level, we consider this percentage to be a valid estimate for the opportunity cost of not receiving a loan.

We assume the opportunity cost as a percentage of the amount of the loan for several reasons: First, the

annual income and the amount of the granted loan are highly correlated for micro-entrepreneurs, as we could corroborate for a similar universe of the private bank's customers. This insight is relevant since the variable "income" is not properly available for all applicants. Secondly, we want to provide a comparison between both costs (monetary loss, which depends on the amount of the loan, and opportunity cost) in terms of the same variable. Finally, it is more intuitive to estimate the profit a good payer may generate (and therefore a social benefit) as a proportion of the money that he/she receives (the amount of the loan), instead of the income he/she already has.

The cost of rejecting a good applicant per cut-off point $p$ is:

$$C_{RG}^{p} = 0.2644 L_p, \ p \in \{0.05, 0.1, ..., 0.95, 1\}, \tag{A.3}$$

where $L_p$ is the average amount of the loans granted for a given cut-off $p$. It is important to notice that for this particular application no interest rate is considered, and only the effect of the inflation is charged for repayment. In a general case, the interest should be subtracted from the profit generated by the loan.

#### A.4.1.3 Cut-Off Point Construction

Given the results of the previous subsections, i.e. knowing the values of $C_{\text{Loss}}^{p}$ and $C_{RG}^{p}$, it is possible to construct the final cut-off point. The optimal value will be simply the one that minimizes the total cost in the test dataset, given by:

$$P_{min} = \text{argmin}_p (C_{\text{Loss}}^{p} \cdot |D(p)| + C_{RG}^{p} \cdot (|G(1)| - |G(p)|)) \tag{A.4}$$

Similarly to equation A.2, $G(p)$ represents the set of customers with probability of default less than $p$, so $|G(1)|$ represents the total number of customers. For equation A.4 we assume a Benthamite welfare function, commonly used for social economics without assuming an uneven distribution of goods, in which an Euro of foregone benefit to society has the same weight as an Euro of loss from default for the financial institution, since the Euro foregone should have been used for a new loan or for improvement of the welfare of a borrower, the damage is indistinct (Arrow and Debreu, 2002).

$P_{min}$ is the cut-off point that minimizes the total cost of misclassification. It should be noted that this

expression is valid only if the value of the objective variable used to estimate the logistic regression model is 1 in case the loan defaults, as is recommended when constructing scorecards (Anderson, 2007).

### A.4.2 Model Follow-Up

As in many other projects, the users of our credit scoring systems asked for a way to update the respective models. This practical need has also been recognized in the literature. According to Thomas et al. (2002), a statistical model has an average lifetime of two years before it begins to lose predictive capacity, which directly translates into losses due to a higher default rate as well as higher provisions. A way to prevent the described loss in predictive capacity would therefore be attractive for any credit-granting company, making model follow-up a challenge and an interesting research opportunity.

The need for follow-up in credit scoring also has a very strong regulation component: credit risk models must be first approved by the national supervisor before first use, and this procedure can be time consuming, expensive (since resources must be allocated), and overall stressing the operations of the credit risk area. These facts block simply updating the model every six months, and incentives the institution to wait as much as possible to change the model, usually when losses are already being perceived. Correct follow-up should help in avoiding these losses from taking place. Additionally, the developing a credit scoring model is not just related to estimating the coefficients, since implementing a new model stresses much more than just the risk analysis department in the company. The costs saved in training, systems implementation time and resources, and other overhead costs that arise when a new model is implemented encourage the use of follow-up techniques over implementing a new model from scratch.

Similar to model construction, model follow-up must be easily understandable by its users and easily applicable. Additionally, it must provide an accurate picture of the population shifts. We applied the following three approaches for model follow-up:

1. Variables' Discriminatory Capacity: Over time, variables can lose their discriminatory capacity. This was measured independently of the respective model, using the same procedures performed for variable selection. We applied K-S or $\chi^2$-tests using the predicted result as the splitting variable on each of the regressors in the model. This way we can measure the discriminatory capacity of the variables on a monthly or weekly basis using currently granted loans, closely following the market movement and providing early alerts if necessary.
2. Discriminatory Capacity of the Model: The model may also lose discriminatory capacity as the market

changes. Unfortunately, this can only be determined exactly once a certain loan has been paid off or – in the opposite case – the respective customer defaults. We applied the following alternatives which can be considered in order to measure the model's performance periodically. We compute a score for each customer who defaulted or paid off the loan each month. Next a confusion matrix is constructed and percentages of false positives and false negatives are calculated. If any of these measures is above a predefined threshold, for example ten percent of the reported test accuracy, then an alert is given. This test allows following the performance of the model closely.

3. Change of Distribution of the Variables: The most challenging test, however, was to detect whether or not a significant change occurred in the distribution of the variables. The underlying question is if a small change in the variables' distribution is a risk for model performance. Any model should allow for certain variations in terms of the estimated coefficients, basically because a statistically correct sample of a population may also represent a slightly different one. Therefore standard K-S and $\chi^2$-tests turned out not to be useful, since they are too sensitive to small variations. We solved this problem by constructing an empirical test as described in the following subsection using the model coefficients and their standard deviations as proxies. If a variable $j$ has an associated estimated coefficient $\hat{\beta}_j$ and an estimated standard deviation $\hat{\sigma}_j$, it follows from the asymptotically normal behaviour of $\beta$ coefficients in a logistic regression model (Hosmer and Lemeshow, 2000), that a 95% interval of confidence for the population parameter $\beta_j$ is:

$$\hat{\beta}_j - 1.96\hat{\sigma}_j \leq \beta_j \leq \hat{\beta}_j + 1.96\hat{\sigma}_j \qquad j = 1,...,N \tag{A.5}$$

The follow-up problem consists of measuring the shift between two different data sets: the original one used to construct the logistic regression model, and a second one with new cases. Formally, we have:

- Original data set $\mathbf{x}$ and estimated parameters $\hat{\beta}_j$ associated with variable $j$, $j = 1,...,N$.
- New data set $\mathbf{x}'$, with new cases $\mathbf{x}'_i = (x'_{i1},\ldots,x'_{iN})$, and observed outputs $y'_i$, $i = 1,\ldots,NC$ where $NC$ is the number of new cases.
- Estimated default probabilities for the new cases $p(\mathbf{x}'_i)$ $(i = 1,...,NC)$ obtained from the original models.

The respective literature provides some approximations to solve this particular problem. The closest model to the one presented here was presented by Zeira et al. (2005). It takes into account a measure of the

shift, not just if a shift has occurred. The authors developed a statistical test for general models considering the output errors, assuming that they distribute normally, and that the variables are identically distributed. A variation of this approach, proposed by Cieslak and Chawla (2007), considers model evaluation in two steps: global discriminatory capacity of the model, and changes in the distribution of the variables instead of measuring them independently. Castermans et al. (2010) recently developed a framework for monitoring and validating the use of risk models within the context of Basel II accord. They consider a stability index to monitor the changes in the distribution of variables and a discrimination index for global performance.

#### A.4.2.1 Statistical Test for Model Follow-Up

The basic idea of the test we developed is to check whether the new coefficients $\hat{\beta}'_j$ estimated from the new dataset $(x'_{i1},\ldots,x'_{iN},y'_i) \in \Re^{(N+1)}$ $(i = 1,...,NC)$ are still within the confidence interval belonging to the original estimators $\hat{\beta}_j$ of the variables.

Since the variables of the new dataset are constructed the same way as those from the original one (a necessary condition for applying the model), estimating the new coefficients $\hat{\beta}'_j$ is straightforward and can be done at a very low computational cost.

With the new estimators $\hat{\beta}'_j$, and their standard deviations $(\hat{\sigma}'_j)$ that are obtained from the new dataset, a new statistic for the population values of the variable $\beta'_j$ can be constructed (Greene, 1993). If the new sample is large enough (necessary condition for estimating the logistic regression parameters), and considering the normality of the coefficients, the following is fulfilled:

$$\frac{\hat{\beta}'_j - \beta_{ref}}{\hat{\sigma}'_j} \rightsquigarrow t_{NC-N}, \tag{A.6}$$

where the statistic has a $t$-distribution with $NC - N$ degrees of freedom. The scalar $\beta_{ref}$ represents the assumption about the population parameter. We constructed a statistical test to verify whether or not the estimated new parameters are still within the confidence intervals obtained for the original dataset considering two one-sided hypothesis tests, as shown in (A.7).

$$\begin{array}{l} H_0 : \hat{\beta}'_j = \beta_{lo} \\ H_1 : \hat{\beta}'_j < \beta_{lo} \end{array} \textit{ and } \begin{array}{l} H_0 : \hat{\beta}'_j = \beta_{up}, \\ H_1 : \hat{\beta}'_j > \beta_{up} \end{array} \tag{A.7}$$

where $\beta_{lo} := \hat{\beta}_j - 1.96\hat{\sigma}_j$ and $\beta_{up} := \hat{\beta}_j + 1.96\hat{\sigma}_j$.

The test checks whether or not the new parameters $\hat{\beta}'_j$ are still within the respective bounds determined by the previous model. In this case, i.e. the null hypothesis $H_0$ is not rejected, the distributions of the variables did not change significantly, which reconfirms the model's stability. To apply this test, the number of cases must be sufficient to ensure normal distribution of the beta coefficients, which should be given if the procedure is conducted every of three to six months.

Our approach differs from the introduced models in several points: first, by being model-dependent (uses the information and variability of the original model), it shows a more business-oriented focus on the concept drift approach. The other methodologies focus on whether a change occurred, or on the absolute strength of this change, not in if the change proposed is relevant for the model, as we believe should be. This fact ties the usefulness of our approach tightly to credit scoring, where logistic regression is widely used and the changes on variables can be much more sudden, in contrast with the other methods that present a more general approach. Another difference, contrasting Hellinger's distance for concept drift (Cieslak and Chawla, 2007) to our method, is that the former is only relevant for categorical variables, whereas our approach is applicable for both continuous and binary variables. The use of K-S and $\chi^2$ measures, also common practice according to the respective literature, is not recommended, since they are known to be very sensitive to small changes in the variables' distributions, and are therefore prone to false alarms. Finally, a difference between our approach and the one by Castermans et al. (2010) is that the bounds on the "danger zone" of the latter are defined focusing on a percentage on the entropy in the comparison of the distributions, which is somewhat arbitrary, whereas our approach is tied to a hard bound on the limits given by the certainty on the original parameter estimation, which we believe is an advantage.

## A.5 Results

In this section we present the results we obtained applying the proposed methodology to the governmental organization we worked with. The methodology employed, however, is generic and has also been used in our projects for the private sector. In particular we analysed two different datasets with loans that ranged

from one to five years duration and amounts that varied between EUR 175 and EUR 17,500. Both datasets present an average granted loan of EUR 1,500.

The first data set (Universe U2) contains 41,200 long-term loans for new customers during a period of 12 years (1996 to 2007), and presents a high number of defaulters, with a default rate of 26.2%. The second dataset (Universe U4) contains 110,000 long-term loans for renewing customers during the same period, and shows a default rate of 17.9%.

The following subsections present results that were obtained in each one of the steps of the proposed methodology as described above, paying special attention to the results achieved by our newly developed steps for cut-off point construction and model follow-up.

### A.5.1 Variable Selection

The repository created for datasets U2 and U4 initially contained more than 100 potential input variables. The goal of variable selection is to identify no more than 10-15 variables for model construction.

Variables commonly used in literature can be divided into three different groups: socio-demographic variables (customer provided, age, income, etc.), internal data (evolution of previous loans, other products, etc.), and external data (outstanding debts, checking accounts, etc.) (Anderson, 2007). All three were present in the original dataset, and all classical indicators of debt evolution were built, if possible. However, since the segment we analysed did not have access to financial services previously, the borrowers did not possess common debts and income variables. This fact made it even more difficult to develop the respective scoring systems than those presented in literature, and made most of "normal" variables useless.

Using simple filter methods for feature selection, we removed highly concentrated variables (i.e. more than 99% of cases that have the same feature value) and obviously irrelevant ones detected by K-S and $\chi^2$-tests, reducing the number of remaining variables to fewer than 100. Subsequently, we applied forward selection and backward elimination for logistic regression obtaining a manageable number of features for each model.

During this variable selection process we maintained very close interaction with our customers, in particular with business experts and future users in order to assure suitable input variables for the respective models. This interaction is of utmost importance, since it adds business knowledge to the selection process and assures model acceptance by the intended users. However, in some cases we obtained surprising re-

sults, such as the case of the income variable, which was expected to be an important input. But as already mentioned, income seems not to be a very good variable when it comes to predicting micro-entrepreneurs' paying behaviour using credit scoring models. Our analyses confirmed this assumption. This is completely different from, for example, the mass consumer segment, in which the income varies much more, and is a relevant variable, especially when constructing indicators associating it with debt.

The final variables selected for U4 are divided into two groups. The socio-economic variables include Economic Activity (Activity), the sector of the economy that the customer is immersed in (through his/her job or company). The large number of activities was clustered to diminish the deviation and to improve interpretation of the variable, bringing the 47 different activities into three homogeneous groups of activities (Activity_A, Activity_B, and Activity_C); the ownership of housing (Ownership), whether the customer owns, rents, or has other types of agreements in his/her current home. Four classes are recognized: Owner, Tenant (Rent), Shares Tenant (Share), or others; the number of productive properties the borrower controls, i.e. properties that are necessary to develop the respective activities and therefore different from ownership, clusterized into zero, one, or more (NumProp_One, Numprop_More), ; a clusterization of the regions in the country (Region_A, Region_B, Region_C). Finally, the age of the customer in years, normalized to $[0,1]$ (Age), or transformed using the natural logarithm (LogAge) is included. It is known that in general age should be treated as a discrete variable to account for non-linearities. However, in our case the range of ages was much more restricted than in normal consumer loans and the behavior was much more linear regarding age. For simplicity reasons we used this justification and treated age as linear variable.

The second group characterize the credit history of the customer, including the number of current or parallel loans (NumCurr), the number of closed loans (NumClosed), the average term for all loans granted previously (AvgTerm), a binary variable indicating whether the customer has been in arrears for any of his/her past loans(PrevArr), the percentage of the paid instalments of previous loans that have been in arrears (PercArr), and the maximum number of days that the customer was in arrears for any previous instalment (MaxArr).

### A.5.2 Model Results

Applying the methodology for credit scoring described above, we obtained a logistic regression model for each one of the two datasets, U2 and U4. Table A.2 displays the respective model parameters. The final models are characterized by 41,200 cases and 10 variables (in the case of U2), and 110,000 cases and 13 variables (in the case of U4). In both instances we used approximately 80% of cases for model construction

Table A.2: Parameters Obtained for Logistic Regression model. New customers (left) and renewing customers (right).

(a) New Customers (U2)

| Variable | β | S.E. | Sig. |
|---|---|---|---|
| Ownership_Owner | -.421 | .044 | .000 |
| Ownership_Let | -.071 | .057 | .216 |
| Ownership_Share | .184 | .147 | .210 |
| LogAge | -.342 | .047 | .000 |
| NumProp_One | .614 | .058 | .000 |
| NumProp_More | .111 | .066 | .092 |
| Activity_A | .320 | .052 | .000 |
| Activity_B | -.022 | .054 | .677 |
| Region_A | .093 | .038 | .015 |
| Region_B | -.569 | .044 | .000 |

(b) Renewing Customers (U4)

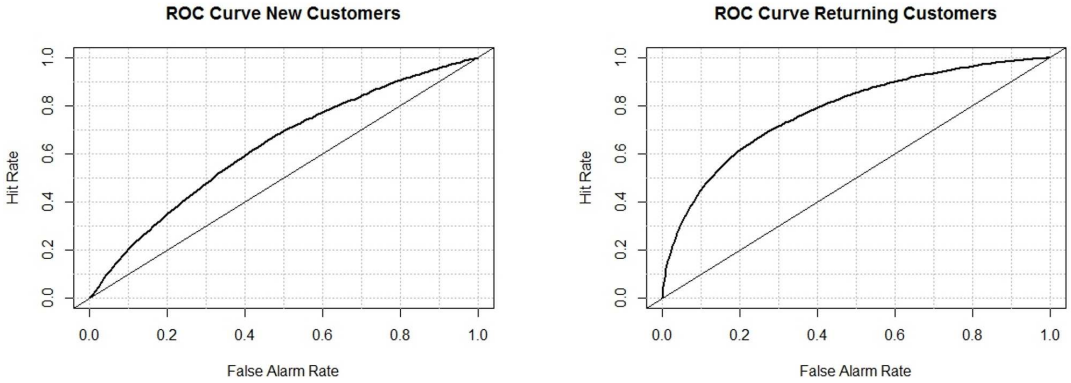
| Variable | β | S.E. | Sig. |
|---|---|---|---|
| Region_A | .092 | .037 | .011 |
| Region_B | -.238 | .032 | .000 |
| Ownership_Owner | -.293 | .037 | .000 |
| Ownership_Let | .076 | .051 | .139 |
| Ownership_Share | .354 | .114 | .002 |
| NumProp_One | .497 | .034 | .000 |
| NumProp_More | .196 | .035 | .000 |
| LogAge | -.436 | .046 | .000 |
| NumClosed | -.090 | .004 | .000 |
| NumCurr | -.034 | .013 | .007 |
| PrevArr | 1.493 | .032 | .000 |
| PercArr | .089 | .018 | .000 |
| MaxArr | .001 | .000 | .000 |

and the remaining 20% for model evaluation.

To evaluate the obtained models, common accuracy measures were estimated for both datasets, with an unsurprising result: it is much more difficult to construct a logistic regression model for new customers than for renewing ones. The evaluation dataset associated with new customers (U2) presented an Area Under the Curve (AUC) of 0.6314, and a K-S maximum distance of 0.1991, while the evaluation dataset associated with renewing customers (U4) presented an AUC of 0.7795 with a K-S maximum distance of 0.4204. This represents a 15% increase in AUC and more than double K-S maximum difference between the two datasets. It is interesting to note that this result is very much in line with consumer credit scoring models, with similar adjustments to those observed in practice. The corresponding ROC curves are shown in Figure A.1.

When studying the accuracy on a case-by-case basis, the models present 74.8% accuracy for non-defaulters, and 65.7% for defaulters from dataset U2, and 78.2% accuracy for non-defaulters, and 76.0% for defaulters from dataset U4, assuming a cut-off point of 0.5. Again, the overall measure accuracy provides better results for known customers (U4) than for new ones (U2). As can be observed from Table A.3 and Table A.4, however, the proposed models identify better among defaulters in U2, a result which is intuitive, since the behaviour of good payers is easily observed during repayment, and this additional information allows for better discrimination of good payers. Furthermore, in our credit scoring projects we observed that good payers in general show a much more homogeneous behaviour than defaulters which present a multiplicity of reasons to default.

Analysing the obtained results reveals that the most important issue for determining paying behaviour is the way a customer handles his/her budget, which is reflected in variables associated with the number of




(a) New Customers (b) Returning Customers

Figure A.1: ROC Curves for the Two Datasets: New Customers (U2, AUC=0.6314) and Renewing Customers (U4, AUC=0.7795).

days in arrears that the customer has.

### A.5.3 Results of Cut-Off Point Construction

As was explained above, for cut-off point construction we need tables that consist of the solicited loans, the estimated default probability, and the loss incurred when defaulting. To determine cut-off points, the calculated default probability is replaced by the closest upper value of the 0.05 intervals used. That is, if the estimated probability is, for example, 0.543, it is replaced by 0.55. Segmenting the cut-off values allows grouping the loans and analysing the effects of changing the cut-off policy over a batch of loans, instead of on a per-loan basis, which greatly simplifies the analysis.

To construct the tables as shown e.g. in Table A.3, we first identify the customers in each 0.05 interval and then calculate accuracy in the respective intervals. This is done separately for Good Customers (column 2 of each table) and for Defaulters (column 3 of each table). The total number of correctly classified cases is presented in column 4; see e.g. Table A.3.

The subsequent columns are used to estimate the related costs of rejecting a good applicant and accepting a bad applicant, respectively. Column 5 presents the average credit amount solicited by good payers who would be rejected if the corresponding cut-off point were used. Column 6 was calculated using Equation (A.2), considering the average observed loss per-customer up to that cut-off point.

Table A.3: Cut-off Table for New Customers (U2). Cut-off points maximizing accuracy (0.95) and minimizing total cost (0.55) are marked as **bold**, respectively. All monetary amounts in EUR thousands.

| Cut-Off | Correctly Classified | | | Avg. | Avg. | Cost | Cost | Total |
|---|---|---|---|---|---|---|---|---|
| | Good | Defaulter | Total | Amount | Loss | (Good) | (Def.) | Cost |
| 0.4 | 5,847 | 9,828 | 15,675 | 967 | 1,789 | 6,279 | 1,731 | 8,010 |
| 0.45 | 8,504 | 9,239 | 17,743 | 900 | 1,614 | 5,208 | 2,513 | 7,722 |
| 0.5 | 11,768 | 8,458 | 20,226 | 831 | 1,402 | 4,093 | 3,278 | 7,371 |
| **0.55** | **15,340** | **7,457** | **22,797** | **784** | **1,226** | **3,120** | **4,092** | **7,213** |
| 0.6 | 18,880 | 6,176 | 25,056 | 740 | 1,102 | 2,254 | 5,092 | 7,346 |
| 0.65 | 22,142 | 4,786 | 26,928 | 696 | 1,009 | 1,519 | 6,062 | 7,581 |
| 0.7 | 25,234 | 3,304 | 28,538 | 644 | 941 | 878 | 7,052 | 7,931 |
| 0.75 | 27,787 | 1,916 | 29,703 | 575 | 886 | 396 | 7,870 | 8,267 |
| 0.8 | 29,404 | 790 | 30,194 | 493 | 848 | 129 | 8,481 | 8,610 |
| 0.85 | 30,180 | 153 | 30,333 | 354 | 820 | 20 | 8,725 | 8,745 |
| 0.9 | 30,376 | 7 | 30,383 | 243 | 813 | 1 | 8,770 | 8,771 |
| **0.95** | **30,396** | **0** | **30,396** | **0** | **812** | **0** | **8,770** | **8,770** |
| **Total** | **30,396** | **10,796** | **41,192** | - | - | - | - | - |

Table A.4: Cut-off Table for Renewing Customers (U4). All monetary amounts in EUR thousands.

| Cut-Off | Correctly Classified | | | Avg. | Avg. | Cost | Cost | Total |
|---|---|---|---|---|---|---|---|---|
| | Good | Defaulter | Total | Amount | Loss | (Good) | (Def.) | Cost |
| 0.4 | 53,805 | 15,563 | 69,368 | 1,666 | 620 | 15,936 | 2,510 | 18,447 |
| 0.45 | 60,374 | 14,473 | 74,847 | 1,711 | 641 | 13,393 | 3,296 | 16,690 |
| 0.5 | 66,129 | 13,428 | 79,557 | 1,740 | 665 | 10,976 | 4,115 | 15,091 |
| 0.55 | 71,090 | 12,292 | 83,382 | 1,765 | 690 | 8,820 | 5,050 | 13,870 |
| 0.6 | 75,243 | 11,093 | 86,336 | 1,788 | 727 | 6,969 | 6,191 | 13,160 |
| 0.65 | 78,890 | 9,738 | 88,628 | 1,811 | 738 | 5,311 | 7,292 | 12,604 |
| **0.7** | **82,004** | **8,312** | **90,316** | **1,876** | **746** | **3,957** | **8,432** | **12,390** |
| 0.75 | 84,642 | 6,685 | 91,327 | 1,980 | 747 | 2,796 | 9,659 | 12,456 |
| **0.8** | **86,737** | **5,054** | **91,791** | **2,118** | **766** | **1,818** | **11,158** | **12,977** |
| 0.85 | 88,418 | 3,358 | 91,776 | 2,295 | 797 | 950 | 12,950 | 13,901 |
| 0.9 | 89,491 | 1,650 | 91,141 | 2,414 | 838 | 315 | 15,056 | 15,371 |
| 0.95 | 89,894 | 578 | 90,472 | 2,107 | 887 | 50 | 16,880 | 16,931 |
| 1 | 89,985 | 0 | 89,985 | 0 | 911 | 0 | 17,861 | 17,861 |
| **Total** | **89,985** | **19,614** | **109,599** | - | - | - | - | - |

Some considerations are relevant: first, collateral is not considered in this segment since the institution would incur in important social and direct costs when trying to recover such items. Consequently, there is no active policy for their collection. Second, the exposure of the loan considers only the amount due (amortization), not the interest collected. This is done to ensure transparency in the reported losses, since interest rates may vary from borrower to borrower.

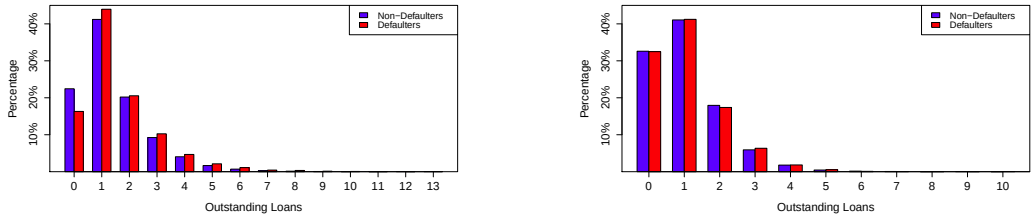
Columns 7 and 8 present, for each possible cut-off point, the cost of rejected loans and the cost of accepted loans, respectively. The cost of rejected loans is obtained by multiplying the cost of rejecting a good applicant ($0.2644 \cdot$ Column 5) and the number of good applicants that would have been rejected for a given cut-off point (from Column 2, all good borrowers -last row- minus the value for the cut-off point). On the other hand, the cost of accepted loans is calculated by multiplying the cost of accepting a bad payer, from Column 6, by the number of bad payers that would have been accepted in that cut-off (from Column 3, all defaulters -last row- minus the value for the cut-off point).

The total cost (Column 9) is finally obtained by adding the cost of rejected loans and the cost of accepted loans (Column 7 + Column 8). Both resulting tables are presented in Table A.3 and Table A.4, respectively.

These tables reveal very interesting insights into the respective business. First, the proposed cut-off points are reasonable in the light of the risk associated with each of the two datasets. For new customers (U2), the total cost is the dominating factor taking into account that it is very difficult to determine the correct class of a first-time customer and that very high default rates have been obtained for this segment. Consequently, the suggestion is a very conservative 0.55 cut-off point, which translates into an acceptance rate (coverage) of only 45%. Evaluating the remaining requests carefully by a committee of experts considering the cut-off point associated with accuracy (0.95) is recommended.

The cut-off points for dataset U4 present a different picture, since for renewing customers much more information is available and the respective segment presents lower default rates. The cut-off point that minimizes total costs is 0.7 which translates into a direct acceptance rate of 85%. Loans with calculated reject probability greater than the cut-off point that maximizes accuracy (0.8) are rejected directly, leaving only a 7% of solicited loans to a committee.

It is also interesting to analyse the expected savings this policy would have generated. The current policy has a cost of 8,770,881 EUR for dataset U2, and 17,861,211 EUR for dataset U4 (considering only costs associated with loss). A conservative calculation can be performed to estimate the savings of using the models, considering that the cut-off point that minimizes costs is used, i.e. there is no committee.




(a) 2000-2004 (b) 2005-2007

Figure A.2: Distribution of variable NumCurr for U4, period 2000-2004 (left) and 2005-2007 (right)

Table A.5: Follow-Up results, new customers (U2).

| **Variable** | $\hat{\beta}'$ | $\hat{\sigma}'$ | $\hat{\beta}$ | $\beta_{lo}$ | $\beta_{up}$ | $t_{lo}$ | $t_{up}$ | $p_{lo}$ | $p_{up}$ |
|---|---|---|---|---|---|---|---|---|---|
| Ownership_Owner | .283 | .096 | .559 | .416 | .701 | -1.378 | -4.347 | .084 | 1.000 |
| Ownership_Let | .216 | .424 | .527 | .021 | 1.033 | .461 | -1.926 | 1.000 | 1.000 |
| Ownership_Share | .167 | .114 | .367 | .216 | .518 | -.426 | -3.083 | .335 | 1.000 |
| LogAge | -.141 | .119 | -.659 | -.821 | -.496 | 5.707 | 2.984 | 1.000 | **.001** |
| NumProp_One | -.295 | .102 | -1.108 | -1.267 | -.948 | 9.512 | 6.394 | 1.000 | **.000** |
| NumProp_More | -.795 | .224 | -1.953 | -2.310 | -1.597 | 6.757 | 3.579 | 1.000 | **.000** |
| Activity_A | .212 | .090 | .169 | .035 | .303 | 1.966 | -1.003 | 1.000 | 1.000 |
| Activity_B | .171 | .101 | -.546 | -.693 | -.400 | 8.519 | 5.632 | 1.000 | **.000** |
| Region_A | -.167 | .093 | -.440 | -.609 | -.270 | 4.782 | 1.117 | 1.000 | .132 |
| Region_B | .015 | .082 | -.100 | -.219 | .019 | 2.860 | -.041 | 1.000 | 1.000 |

Under this setting, the savings for our governmental institution, and therefore for society, when using the models in dataset U2 are $8,770,881 - 7,213,219 = 1,557,662$ EUR, which is equivalent to 4.2 EUR per loan granted. The savings in U4 are $17,861,211 - 12,390,124 = 5,471,087$, or a staggering 49.9 EUR per loan.

There is no need to emphasize the usefulness of the proposed methodology and of the use of credit scoring models in general, considering the substantial savings that are produced by the use of risk assessment models.

### A.5.4 Follow-Up Results

In order to study the performance of the proposed follow-up methodology, we constructed models for datasets U2 and U4 using the loans granted during the period 2000-2004, and then we analysed the models' changes for the period 2005-2007. We divided the sample into an 80 percent training set for both periods, and a test set consisting of 20 percent of the data, constructing two test sets and two training sets. We trained

four models, one for each period, and for each dataset (U2-U4). Figure A.2 shows an example of how a particular variable (number of current loans from U4) changes its distribution in a period of three years, affecting its influence in the model.

To motivate the effects of concept drift we estimated the loss in discriminatory capacity applying the model with "old" data to the new test set: the AUC of the model built with new data corresponds to 0.6404, and to 0.5924 using the original model, representing a drop of eight percent in AUC. Finally, from the work of Blochlinger and Leippold (2006) it can be deduced that a drop of 0.01 in the K-S statistic can translate in a loss of up to two percent in the net utility of the lender, and in this case the statistic goes down from 0.2359 to 0.1622, representing 75 base points less, or up to 15 percent loss of utility for the lender, which can have catastrophic effects. We believe this reflects the strong need for proper model follow-up.

The results obtained for datasets U2 and U4 using the test proposed in Equation (A.7) are shown in Table A.5 and Table A.6, respectively. For each variable we computed the coefficients $\hat{\beta}'$ for the period 2005-2007, its standard deviation $\hat{\sigma}'$, the coefficients $\hat{\beta}$ for period 2000-2004, its lower bound $\beta_{lo}$ and upper bound $\beta_{up}$, the t-statistics $t_{lo}$ and $t_{up}$ for both hypothesis tests presented in Equation A.7 and the one-sided P values $p_{lo}$ and $p_{up}$. These latter values can be interpreted as the probability that the statistic $\hat{\beta}'$ would differ as much as the boundaries obtained from the original dataset in the direction specified by the hypothesis just by chance, even though the parameter is actually within these boundaries (assuming that the null hypothesis is true). P values below 0.05 can be considered low enough to affirm that the new statistic differs significantly from the original one.

It can be concluded from these experiments that the models present an important loss in performance, especially regarding prediction of non-defaulters. One of the main reasons for this change can be seen in the variables, where four out of ten present critical changes in their distribution for new customers (Age, Number of Properties in two levels, and Activity), affecting the performance of the model. The change is even greater for the universe of renewing customers, where nine out of thirteen variables present significant changes in comparison with the original data set. Furthermore, many of the affected variables become irrelevant in the new data set, which is a clear sign that the models need to be re-adjusted. These results underline the potential of the model follow-up methodology proposed in this paper.

Table A.6: Follow-Up results, renewing customers (U4).

| **Variable** | $\hat{\beta}'$ | $\tilde{\sigma}'$ | $\hat{\beta}$ | $\beta_{lo}$ | $\beta_{up}$ | $t_{lo}$ | $t_{up}$ | $p_{lo}$ | $p_{up}$ |
|---|---|---|---|---|---|---|---|---|---|
| Region_A | .112 | .078 | .121 | -.005 | .247 | 1.492 | -1.725 | **.000** | 1.000 |
| Region_B | -.197 | .075 | -.135 | -.273 | .003 | 1.010 | -2.649 | 1.000 | 1.000 |
| Ownership_Owner | .268 | .094 | .689 | .530 | .848 | -2.775 | -6.154 | **.003** | 1.000 |
| Ownership_Let | .379 | .293 | .919 | .494 | 1.345 | -.391 | -3.294 | .348 | 1.000 |
| Ownership_Share | .122 | .095 | .632 | .488 | .775 | -3.852 | -6.874 | **.000** | 1.000 |
| NumProp_One | .086 | .074 | -.687 | -.808 | -.566 | 12.045 | 8.790 | 1.000 | **.000** |
| NumProp_More | -.018 | .084 | -1.143 | -1.286 | -1.000 | 15.047 | 11.654 | 1.000 | **.000** |
| LogAge | -.397 | .126 | -.636 | -.814 | -.458 | 3.309 | .488 | 1.000 | .313 |
| NumClosed | -.067 | .007 | -.114 | -.131 | -.096 | 9.536 | 4.307 | 1.000 | **.000** |
| NumCurr | -.038 | .035 | .104 | .057 | .152 | -2.659 | -5.358 | **.004** | 1.000 |
| PrevArr | 1.351 | .093 | 1.877 | 1.751 | 2.004 | -4.322 | -7.051 | **.000** | 1.000 |
| PercArr | -.023 | .026 | .077 | -.001 | .154 | -.837 | -6.702 | .201 | 1.000 |
| MaxArr | .001 | .000 | .002 | .002 | .002 | -3.832 | -7.236 | **.000** | 1.000 |

## A.6 Conclusions and Future Work

Granting loans to micro-entrepreneurs is a very important business opportunity in developing countries. As a country develops, granting such loans is slowly moving away from public institutions and is being considered a real business opportunity for private organizations, such as banks. The high risk, however, associated with micro-entrepreneurs is one of the main problems that hinders the expansion of this type of loans, and consequently a faster development of the respective economies. This explains the utmost importance of adequate risk assessment models that are tailored to the particular characteristics of micro-entrepreneurs.

In this paper, we have described the successful adaptations of the standard KDD process to the particular needs of public and private financial institutions that grant loans to micro-entrepreneurs, using logistic regression as the classification method. This procedure provides good results and a useful interpretation of the respective input variables.

Traditional credit scoring variables such as “income” were not relevant. Micro-entrepreneurs have similar incomes and therefore this variable does not discriminate. We also tried related variables using income, such as income/debt, etc. which neither discriminated well.

Another important conclusion is to use variables that can be corroborated. Ambiguous variables describing characteristics that cannot be proven easily are not useful since customers and/or salespersons know how to answer certain questions in order to improve their chances of getting the loan.

The proposed cut-off point methodology explicitly quantifies the lost benefit for society when a loan for a good customer is not granted. Our numerical experiments underline the potential impact such a methodology

might generate, and help in quantifying the benefit of using statistical models in practice, with important savings when compared to the absence of such models.

Micro-entrepreneurs represent a very volatile market, and are therefore very sensitive to changes in economic conditions, which makes the nature of the respective credit granting operations very dynamic and subject to constant change. This is reflected by shifting risk factors. Tools that detect such shifts are very attractive for practitioners but not always available in standard solutions for credit scoring. In this paper we introduced statistical tests for model follow-up that were developed in our projects and provided excellent results.

By using adequate methodologies for credit risk management, the market of loans for micro-entrepreneurs will continue to grow as more private companies will offer such loans; see e.g. Kim and Sohn (2007) and Kim and Sohn (2010) for a similar situation. This, in turn, will foster their capacity to innovate and generate growth. OR-methodologies will contribute to a sustainable development of the respective countries as has been shown already for many other cases in White et al. (2011).

# Acknowledgement

The first author acknowledges CONICYT for the grants that supports this work (AT-24110006, NAC-DOC: 21090573). All authors acknowledge the support of the institution which provided the data. The work reported in this paper has been partially funded by the Institute of Complex Engineering Systems (ICM: P-05-004-F, CONICYT: FBO16) and the Finance Center of the Department of Industrial Engineering, Universidad de Chile, with the support of bank Bci.