A Credit Risk Management Model for a Portfolio of Low-income Consumer Loans in Mexico

By

Jorge Alberto Jiménez Montesinos

M.B. A.
Instituto de Empresa Business School, 2004

SUBMITTED TO THE MIT SLOAN SCHOOL OF MANAGEMENT IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN MANAGEMENT STUDIES AT THE MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2014

© 2014 Jorge A. Jiménez Montesinos. All Rights Reserved.

The author hereby grants to MIT permission to reproduce and to distribute publicly paper and electronic copies of this thesis document in whole or in part in any medium now known or hereafter created.

Signature of Author: ____________________________

MIT Sloan School of Management

May 9, 2014

Certified By: ____________________________

Mark P. Kritzman
Senior Lecturer in Finance at MIT Sloan School of Management
Thesis supervisor

Accepted By: ____________________________

Michael A. Cusumano
SMR Distinguished Professor of Management
Faculty Director, M.S. in Management Studies Program
MIT Sloan School of Management
A Credit Risk Management Model for a Portfolio of Low-income Consumer Loans in Mexico

By

Jorge Alberto Jiménez Montesinos

Submitted to the MIT Sloan School of Management on May 9, 2014 in partial fulfillment of the requirements for the degree of Master of Science in Management Studies

ABSTRACT

Low-income consumer lending (LICL) in Latin America has experienced a boom in recent years. This has attracted the interest of a large number of financial players eager to capture a portion in this still under-banked segment. Despite this huge market potential, credit risk management in this segment is still mainly based on the subjective expertise of credit managers, with few exceptions making use of robust statistical techniques. The result of this subjective decision process is a sub-optimal concession of loans that leaves this strategic segment without adequate financing opportunities. In this work we develop a cutting-edge probability of default (PD) model specifically designed for the LICL segment. This research is one of the first academic works that explores the applicability of four cutting-edge quantitative methods with real operation data from one of the pioneers in the low-end credit segment in Mexico: Mimoni Group. The analysis was carried out on a sample of 2,000 loans including 741 defaults and 1,259 non-defaults, spanning the 2013 to 2014 time period. We run a total of 108 models utilizing Logistic regressions, CART models, Random Trees, and Clustering over training and out-of-sample data sets. Our results not only generated powerful models in terms of statistical accuracy in out-of-sample data sets, but also provided a detail list of robust PD predictors (at 95% levels) and their dynamics in explaining the default event for two Mexican low-income customer segments. Our results demonstrate the direct applicability that robust quantitative models have in improving and complementing the lending decision process in the growing LICL segment.

Thesis supervisor: Mark P. Kritzman
Title: Senior Lecturer in Finance at MIT Sloan School of Management
ACKNOWLEDGEMENTS

First and foremost, I thank Hashem for the opportunity to be in one of the most intellectually-enriching learning communities in the world: MIT.

Additionally, I want to express my gratitude to my Thesis Supervisor, Prof. Mark P. Kritzman, for the great support provided during these months. Despite his multiple occupations as the eminent leader he is the Finance industry, he always provided me time to share his invaluable expertise in the credit risk modeling field. Prof. Kritzman not only shared his Finance knowledge with me, but his high caliber as a human being. Thank you Professor for being an example for all Sloanies.

Secondly, I want to thank Prof. Roger Stein and Prof. Alexander Reyngold without whom the present research could have not being possible. I feel thankful to have learned directly from one the most prolific authors in the credit risk management discipline. Prof. Stein's teachings instilled in me a profound passion for the risk modelling discipline that certainly will accompany me in the years to come. On other hand, I do not find words to thank Prof. Reyngold's valuable time to solve my questions and help me understand the economics behind the numbers. His advice was fundamental in the construction of the models included in this research.

Thirdly, I want to thank Prof. Michael Chu for his invaluable guidance and inspiration about how financial innovation can and should be used in the development of segments at the base of the pyramid. I really feel honored having the chance to discuss my ideas with the top thought leader in the Social Enterprise, Microfinance, and Financial Inclusion fields. I also want to thank him for having opened me the door to Grupo Gentera, the Center for Financial Inclusion, IGNIA, and Mimoni Group.

Additionally, I extend my gratitude to Mr. Pedro Zayas, founder and CIO of Mimoni group, whose support was the key piece in the completion of this work. Mr. Zayas not only provided one of the largest and high-quality data bases for low-income consumer lending, but was always open to share his vast experience and inestimable insights as a pioneer in the industry. Mr. Zayas is the clearest example of innovative and social responsible leaders that come out from MIT.

Furthermore, I am exceptionally thankful to Julia Sargeaunt, Chanh Phan, and Christine Bolzan for making the MSMS the life-changing experience that it is. During this year, I felt not only as another master student but as a true member of an extended family. Your advices and cherished friendship will always be present in my heart.

Finally, since I was in high school I always dreamed about studying at MIT, the cradle of innovation and place where mind, passion, and action converge for the advancement of humankind. Today this dream is a reality, which could not have been possible without the support of my parents Marco A. Jiménez Aguilar and Alba Montesinos. You are my strength and source of inspiration. You taught me that dreams come true, and that the best dreams are those that look the betterment of others. I will keep growing to honor your teachings and moral standards.
# Table of Contents

## Chapter 1. Introduction

1.1. Context: Low-income consumer lending in Latin America .......................................................... 12
1.2. Success factors for Latin America low-income consumer lending ........................................... 15

## Chapter 2. Literature Review ................................................................. 17

2.1. Current Credit Risk Toolbox and models. .................................................................................. 20
2.1.1. The Credit Scoring Tools........................................................................................................ 21
2.2. Techniques used in the measurement of credit risk ................................................................. 30
2.2.1. Econometric Models, Classification and Regression Trees, and Random Forests ............. 35
2.3. Statistical Model Validation Techniques and Evaluation Criteria ......................................... 42
2.4. Transforming Statistical insight into Economic terms ............................................................ 71

## Chapter 3. Model Development and Results .................................................. 78

3.1. Data Provide: MIMONI ........................................................................................................... 78
3.2. Database description ................................................................................................................. 80
3.3. Data set cleaning, and variable transformations ......................................................................... 86
3.4. Regressions, CARTS, Random Forests, and “Clustering-then-Predict” approach. ............... 95
3.4.1. Direct approach: Model’s Predictions over whole data set................................................. 97
3.4.2. “Clustering-then-Predict” (CTP) approach: tailoring regressions to specific groups......... 111
3.2 Database description ................................................................................................................. 80

## Chapter 4. Conclusions ............................................................................. 127

## Appendix ................................................................................................. 134

## Bibliography .................................................................................................. 141
"Lasting peace cannot be achieved unless large population groups find ways in which to break out of poverty. Micro credit is one such means. Development from below also serves to advance democracy and human right"

Muhammad Yunus (2006)

Low-income consumer lending (LICL) in many countries in Latin America has advanced steadily at double-digit growth rates in the past five years. This accelerating trend in the consumer credit persists even in nations experiencing financial and economic downturns. In countries like Mexico, Brazil, Colombia, Chile and Peru the growth coincides with rising incomes of many households at the lower rungs of the social pyramid. For example, in 2010, the best performing companies focusing in low-income credits in Mexico and Chile reached ROEs of 42% with annual portfolio growth rates around 20%. This expanding consumer base has attracted a large number of financial players eager to capture market share in this still underserved segment.

However, despite its importance, low-income consumer credit in Latin America is still distant from reaching the state of development and penetration of high and middle income segments. For instance, in Brazil (the largest consumer lending market), low-income banking has only reached 31% penetration rates vis-à-vis an almost 80% rate at upper-income segments, a difference that is even larger in most countries of the region1. Among the reasons behind this lagged development we find that LICL involves paradigms very different than those at the top of the pyramid. Latin America’s traditional retail banking infrastructure is just not appropriate to

---

1 Instituto Brasileiro de Geografia e Estadística; Pesquisa Nacional por Amostra de Domicílios (PNAD) 2012; http://www.ibge.gov.br/home/estatistica/populacao/trabalhoerendimento/pnad2012
secure a sizable part of the market. For instance, under the typical retail bank format, the average acquisition cost per client is around $130, but only annual expected margins of $120 per low-income customer in the best scenario. This is due to the high distribution costs of traditional retail branch involving strong client service and branches maintenance. Therefore, we can see that, to effectively serve this segments, financial institutions need to develop LICL tailored commercial strategies and low-cost distribution infrastructures.

On other hand, although the outlook of the LICL segment is highly promising and quantitative credit risk management has simply became an integral component in “mainstream” financial areas, in low-income segments, lending practice still rely on human-judgment, and in the best cases make just a modest use sophisticated statistical methods. The results of this lag in credit risk management in LICL and over-dependence in subjective process are sub-optimal concession of credits and inflated interest rates.

As Chu (2004) states: “loan sharks charge interest rates that may be anything from 5 percent a day to 5 percent a week...at this rates the increase in productivity goes to the source of capital instead of the entrepreneur” This situation is not sustainable neither for the loan receiver nor for the lender. Then why does capital providers keep demanding such inadequate rates? The answer is the lack of exploitable information and robust analytical models to discriminate accurately, consistently and cost-effectively between defaulting and non-defaulting credits. Among the main reasons why LICL credit risk assessment has remained in the traditional subjective, human-based format we have two closely related causes: 1) a lack of structured default data bases, and 2) and extended skepticism due to an absence of research on the applicability and size of benefits that could be achieved by using sophisticated models.
The purpose of the present research is precisely to provide a concrete proof that the use of combined quantitative methods (as opposed to the prevalent practice of using only expert-based decision or “standalone” algorithms) has direct application for the nascent low-income consumer lending industry in Latin America. We demonstrate this by utilizing real operative information from one of the pioneers in the industry, Mimoni Group, to develop a probability of default (PD) model specifically designed for the LICL segment. While there have been many successful models developed for corporate distress prediction objectives and several studies have been used by academics and practitioners for SMEs², this research is one of the first academic works that explores the applicability of four cutting-edge quantitative methods with real data in Latin America low-end credit segment. The analysis was carried out on a sample of 2,000 loans including 741 defaults and 1,259 non-defaults, spanning the 2013 to 2014 time period. We run a total of 108 models utilizing Logistic regressions, CART models, Random Trees, and clustering over training and out-of-sample data sets.

In the opinion of the author, LICL stands at a historical point were leaders of the industry can exactly replicate the similar development that the credit card and high-income segments in developed countries achieved three decades ago based on structured data and quantitative models. In that case, literally, hundreds of thousands of bad credit card loans and millions of exposures allowed for confident inference ergo to the development of statistical models. This combination of data and quantitative tools underlie the recent evolution of the high-end consumer lending industry. For example, the implementation of transparent and verifiable measures of risk -

independent on sole subjective human criteria- provides investors with objective and comparable information across credits, which permits the involvement of other financial actors through more sophisticated formats such as the increasingly widespread securitization of consumer debt in United States, for example. This type of opportunities are possible not only due to data volume alone but also to the capacity of lenders to process this information in consistent, accurate and cost-effective way through statistical models. The question now is: Can the success of credit modeling in upper-levels of commercial lending be replicated for low-income segments? The current size of Latin American LICL market suggests it is possible.

One of the objectives of this research is to demonstrate the value of structured data bases and the advantages derived from using quantitative models using this input. It is important to mention that we are convinced that the strengths of statistical models are not limited to improved accuracy and power of prediction to discriminate between bad and good credits (optimal cutoffs), but also that such statistical insight can be converted immediately in practical guides and tangible economic returns for lenders in terms of streamlined, automatized decision-making processes for large numbers of credits, fair interest rates, and capital requirements that do not predate low-income borrowers while providing still commercially attractive returns for the providers of capital.

As we mentioned above, the exponential growth of low-income consumer lending inherently implies an unsustainable future for the traditional one-to-one, human loan evaluation and supports the superiority of mechanisms allowing accurate and reliable processing of large batches of credit requests in a consistent and cost efficient way. This is the core of our research and in the coming chapters we will demonstrate our hypothesis through the development of a
Credit Risk model specifically designed for one of the largest LICL markets in the region: Mexico.

The following work is structured as follows. In the remaining part of Chapter 1, we will provide a general context of Latin American low-income consumer lending segment. In Chapter 2, we will present a literature review of the most important ideas and methodologies used today in the credit risk management industry. Among the most important topics in this section we included a detail list of today’s credit risk toolbox providing further detail on the statistical methods used in this work (i.e. Logistic Regressions, Classification and Regression Trees, and Random Forests). Then we discuss important considerations on PD model power, calibration, and relevant validation techniques to avoid overfitting. Finally, we include some pricing techniques to calculate optimal lending cutoffs, and pricing techniques based on a model’s Receiver Operator Characteristic (ROC) curve. In Chapter 3, we provide a description of the data provider (i.e. Mimoni Group) and its credit application process. Then we present the composition of our data base with a detailed description of the 34 variables included in the study, the data cleaning and preparation techniques utilized, and some variable transformations implemented. Afterwards, we run some multicollinearity tests, identifying and eliminating highly correlated variables, and describe the k-means methodology used to identify the two groups used in cluster-specific modeling. Finally, we present our results using: 1) “direct” modeling over the whole data set, and 2) a “Clustering-then-Predict” (CTP) to develop group-specific models. Chapter 5 provides our conclusions and some extensions for further research.
1.1. Context: Low-income consumer lending in Latin America

"Until now business was defined by the goods and services we provided to two billion people—people in the first world and the elites of the world. But today we are on the verge of understanding how to unlock goods and services for the remaining four billion."


In 1970, when inclusive financial strategies as microfinance got its starts, the prevailing view was that to only way to help people at the base of the pyramid was by providing subsidies. Any consideration about charging interest to these low-income segments was severely criticized. But today, doing business in the lower rungs of the social pyramid is beginning to acquire a totally different outlook.

The combination of a significantly large latent market in Latin America and the new attitude toward social enterprise where capital providers recognize they can do good by doing well is changing the LICL completely. For any activity to be part of business, it must be first sustainable, it must be sizable enough, and able to generate enough revenue stream to cover costs and provide a gain. Three factors support the idea that this growth should continue: 1) the wide gap in penetration rates in low-income segments versus high-end lending (see above for an example of Brazil economy); 2) the high ROEs above 40% levels in various countries in the region, and 3) the two-digit annual portfolio growth rate in companies already specializing in lending at the lower rungs of the pyramid.
However, although the overall low-income landscape seems promising, financial intermediaries must recognize that different Latin American countries present different levels of development. We can characterize two main groups based on their level of sophistication, although it is important to mention that we can find a co-existence of different levels within a single country. In the first type, we find large low-income proportions of the population with virtually no access to formal credit. Even when some financial institutions might offer so-called low-income loans, proof-of-income and/or formal job requirements preclude these individuals from qualifying for a loan as many of them work in informal sectors of the economy. Therefore, most financing comes from family and in most cases from local moneylenders known as agiotistas. In these type of markets, around 75% of the low-income population with almost 50% resorting to informal lenders at exorbitant interest rates.

Differing from country to country, interest rates charged for the poor can range from 5% a week to 30% a month, although cases above 150% have been recorded. This figures translate in many cases into annual rates of more than 2,500%, for loans about 10% the size of an average personal loan at a retail bank. Despite the considerable growth of the low-income consumer lending industry in Latin America, the market still belongs to the agiotistas. The reasons behind the low penetration go from incorrect adaptations of the retail distribution model to low-income segments, or inadequate interest rate calculations due to an overdependence on the subjective expertise of credit managers (without extensive support of quantitative tools to support their decision process as we explored in the present research).

In the second type of market, we find basic financial products offering coupled with increasing market penetration. In this markets, financial providers offer secured and unsecured personal loans, point-of-sale financing, and loans from retailers, among others. Example of profitable
formats in these markets include: networks of financial stores offering salary advances and unsecured personal loans, or partnerships with retail chains to finance in store purchases. The formats have evolved differently across countries. In countries such as Chile, retailers dominate the scene having established their own consumer finance banks, while in Brazil traditional banks control the market having absorbed almost all independent low-income lenders. Mexico present a hybrid between these two extremes.

1.2. Success factors for Latin America low-income consumer lending

Within the credit risk management industry the three main challenges for any financial institution are: 1) to concede loans to good credit subjects; 2) to manage exposures effectively; and 3) to secure maximum recovery rates over delinquent loans. However, in Latin America, we have additional challenges that need to be overcome if financial players pretend to serve low-income segments and remain profitable.

First, low-income bankers need conduct the above mentioned risk assessment procedures with unstructured and scarce data. Several pioneers in the field have learned how to combine credit bureau information and other in addition to other sources with intense interaction with clients (e.g. partnerships with retailers) to compensate for scarce information. On the other hand, practitioners have started to explore models tailored by income, product and other demographic variables with acceptable results in discriminating between defaulting and non-defaulting obligors. The present research is one of the first academic works exploring this “Clustering-then-Predict” approach through a combined implementation of Econometric Models (the predominant technique in the industry), CARTs and Random Forests.
Second, highly effective collection procedures play a preponderant role when lending to low-income clients. In this type of market, average expected delinquent rates can reach 40% levels in the first year. However, companies with effective collections operations in place are known to reduce these losses in about twenty percent. This highlights the significant contribution that effective collection has on profitability. Examples of successful collection practices include: preliminary classification of clients by willingness to pay (based on socio-economic algorithms as that of Mimoni Group in Mexico), and collection process managed via centralized facilities to gain economies of scale (even outsourcing or establishing partnership with other financial operators if justified).

Third, given the high volatility of the region, delinquency rates fluctuate constantly thus credit approval rates have to be constantly monitored and calibrated. Further volatility is added as borrowers in the segment are still adapting to a broader menu of products and larger credit lines. This demonstrates that to be effective in the LICL segment it is not enough to have powerful models but also to establish rigorous and continuous calibration processes to prevent unsustainable delinquency rates during bad credit episodes.

Finally, providers of financial services for low-income segments must always keep in mind a low-cost focus. As we mentioned above, traditional retail banking infrastructure is just not appropriate to secure a profitable piece in this growing market. To succeed in this market financial players need to learn how to combine the advantages of scale effective credit processing without losing precision in segmenting clients with different probabilities of default. Mimoni Group, our data provider, is a pioneer using on-line credit application at a mass scale, while obtaining the accuracy and precision required given the segments high volatility, need for constant calibration while providing affordable interest rates. The present research builds on
these same principles demonstrating the value that multisource, structured information coupled with cutting-edge statistical models can have in accessing scale advantages, while maintaining the precision and detail to define customer-specific lending strategies.

In summary, the low-income consumer lending segment represents a huge business opportunity for financial intermediaries in Latin America. Countries like Mexico and Brazil are clear examples that lending at the base-of-the-pyramid is attractive not only for its enormous market potential, but also because it promotes the economic development of these segments. However, to be successful, financial providers need to excel in: 1) designing commercial and distribution models appropriate for these customer economics; 2) implementing highly-effective collection units; and 3) developing scale-efficient credit processes integrating both the accuracy and precision of robust quantitative models and expertise of credit managers.
CHAPTER 2. LITERATURE REVIEW

“All epistemological value of the theory of probability is based on this: that large scale random phenomena in their collective action create strict, non-random regularity”

Gnedenko and Kolmogorov (1954)

Since ancient times, humankind’s main objective has been to understand an uncertain reality using the information it has at hand to minimize its risks. Our epistemological efforts, including socio-economic theories, religions and science, are all efforts to predict and manage our evolving randomness. Randomness is consequence of our imperfect understanding of reality and the multiple forces and relationships behind an event. For almost over a century now. We have turned from deterministic attempts to explain randomness to a new form of predictability that uses this vastness and complexity into our favor. Statistics is our modern way to decipher the patterns behind the chaos through the collection, analysis and modeling of data and with the ultimate objective of inferring the behavior of the collective from the behavior of a simpler, more manageable and representative unit of analysis, the sample.

As stated by Bouchard and Potters (2011), statistics is especially useful when analyzing human aggregates. Financial markets, specifically, are excellent examples of large groups of individuals where the multiplicity of interactions and conflicting interests derive in apparently unpredictable behavior, but whose collective dynamics demonstrate certain regularity and stability across the time. Is in this kind of collectively stationary phenomena where statistics and quantitative modeling can be most helpful.

In the area of credit risk quantification and default prediction, this capacity to draw inferences of future loan exposure through statistical models has taken a preponderant relevance more than
ever. The dramatic credit crisis of 2007-2008 and gigantic losses arising from the exponential use of credit derivatives have made evident the vulnerability of current risk management practices and the need for more sophisticated and accurate tools to measure, monitor and control credit risk. This recent events have just transformed the way credit markets operate.

For instance, until the last decade, most of credit evaluation consisted on a combination of qualitative criteria to evaluate client’s potential credit worthiness coupled with a set of financial ratios to provide a sense of numerical substance to their decisions, a sort of financial justification for regulators. But today, the extreme scrutiny to which financial intermediaries are subject has made thinks drastically different, making both qualitative and quantitative analysis a must in any credit decision process. Concepts such as portfolio volatility, default probabilities and credit exposure correlations, among others, are just part and parcel of today’s normal operative reports at any major financial institution. Today it is a rarity to see a credit manager who does not rely on a robust set of analytical models, rigorous metrics and techniques to evaluate the overall risk to which his company is exposed. Quantitative credit risk management has simply became essential for any company willing to survive in the highly volatile financial environment in which we are living.

The following chapter reviews some of the most important ideas and methodologies regarding credit risk management. In the first section, we will present the main components in today’s credit risk toolbox as well as most important models used in the commercial lending industry. Our discussion of failure prediction models will be focused on econometric techniques, concentrating on discrete-choice models, hazard-rate (duration) models and classification and regression decision trees (C&RT) to predict the probability of default (PD) in commercial loans. This part will include a description of the main variables used by top practitioners to develop
their default prediction models. The second section includes important considerations for PD model validation including measures of model power, model calibration, and techniques to avoid model overfitting. Finally, in the third part, we will discuss the use of optimization and other quantitative techniques to determine the economic value of a model and how PD statistical insight can be used to support more informed lending decision processes (e.g. optimal interest rate, pricing policies and cost-benefit structures). Although many of these techniques have been explored separately –although not extensively- in low-income consumer lending models, there is no work incorporating these methods in a consistent, integrating framework. This is one of the main contributions of this work. As mentioned in past chapters, through an integrated model we will demonstrate that the strengths of statistical approach are not limited to improved accuracy but that can be translated directly into tangible economic and operative returns for financial intermediaries and fair lending costs for borrowers at the base of the pyramid.

2.1. Current Credit Risk Toolbox and models.

This section reviews some of the most important models and techniques used today in the credit risk management field. After first analyzing the most popular quantitative lending tools available in the industry, we will switch to a discussion of both empirical and theoretical default prediction models focusing on the econometric techniques that will use in Chapter 3 as the basis of our model. In statistics, model superiority can only be determined in comparison to other feasible alternatives given that “goodness” is defined in terms of its statistical significance relative to other models. The literature on the field is substantial and the overview presented here is by no
means exhaustive, but we believe it will provide enough context to understand the state of the art, the choice of our model and its strengths relative to other alternatives.

2.1.1. The Credit Scoring Tools

In the following lines we begin our discussion of the main credit scoring tools and the situations in which they are employed.

Consumer Bureau Scores

Credit scoring is a statistical method aimed at predicting the relative likelihood that a loan applicant will default or become delinquent on a credit obligation. These credit scores are based primarily on credit report data provided by Consumer Reporting Agencies (CRAs). Creditors rely extensively in such scores, using them to determine who qualifies for a credit, the interest rates applying, as well as credit limits for each individual. For instance, an applicant with a credit score below 550 is not candidate to receive a prime mortgage, thus will have to go for a subprime mortgage, which normally impose a higher interest rate.

There are various methods to calculate credit scores. The most widely used is the FICO score, which was produced by Fair Isaac Corporation in the 1950s, and which is a number ranging between 300 and 850- he higher the number, the more credit worth the individual is. However, other consumer reporting agencies and even large lenders generate proprietary generic score models (e.g. Equifax's Credit Score, Experian's PLUS score, and TransUnion's CS).

Credit bureau scores have become the gold standard of quantitative lending tools not so much for its accuracy or predictive power, but for its capacity to offer standardized, consistent and cost efficient credit information. Credit scores became widely available in the 1980's. Before this,
human evaluation was the only factor in deciding who received a loan. Credit managers used their past experience in past loan concessions as basis for evaluating new applicants. This process was not only time consuming but was unreliable because of varying interpretation from analysis to analysis.

Credit granting, however, took a huge leap forward when literally millions of loan exposures and hundred thousands of bad credits were pooled together allowing a systematic exploitation of consumer’s credit information in a cost efficient manner. Today, according to Moody’s: “a credit score that captures at least 90% of the measurable risk inherent in a consumer relationship can be purchased for a few dollars... [This], in turn, had allowed the segregation of pools of consumers whose expected loss varies by as much as 10% of notional balances”.

In the United States, this precision, transparency and validated risk measures are the elements that had permitted the securitization of consumer debt, by bundling and offering portfolios customized to different investor’s risk appetites and profiles, increasing market liquidity and reducing interest costs for customers.

In the industry, bureau scoring is not limited only to consumer loans but has application for small businesses. As stated by Berger and Frame (2005), this have acquired recent relevance as analysts realized that “…particularly for the very smallest credits, credit information for owner explains a significant amount of the variation in the performance of small business credits. This may reflect in part a correlation between personal and business success and in part a commingling of the finances of the business and the owner.” On other hand, besides consumer credit scores, some CRAs (e.g. Experian, and Dun & Bradstreet) provide similar scorings to evaluate the risk in extending loans to businesses, underwriting insurance, or investing on a
certain company. However, the most important use of these business score reports is to estimate the credit risk of purchasers of merchandise on credit terms, regardless of the credit quality of the principal owner(s) or firm’s reported financial statements.

Credit ratings

Rating agencies, such as Moody’s, S&P, and Fitch, specialize in assessing the creditworthiness of debt securities issued by different obligors and the credit capacity of the issuers themselves. The ratings produced by these agencies are opinions based on qualitatively and quantitatively performance of the entity, and are estimations on the loss given default and default probability; therefore, they act as default prediction and exposure models to some extent.

These ratings are broadly accepted by the financial community, and the agencies are tremendously influential around the world covering approximately $34 trillion in securities. As stated by Caouette and Altman (2008), “the relevance of these agencies have reached unprecedented levels in US as capital markets have replaced banks as principal source of debt... and agencies have adopted enormous in the management of credit risk.”

The three principal agencies employ different ratings but their scales are somehow comparable. A summary of the rating description for the main agencies is presented below in Table 2.1. The bonds classified with the highest rating (e.g. Moody’s Aaa) have practically no probability of defaulting, with its capacity of repayment deteriorating as rating decreases. It is noteworthy to mention that only bonds with ratings of Baa/BBB or above are classified to be investment grade; below this threshold securities and obligors are categorized as speculative grade (or “junk”
bonds). To provide finer detail, the three main agencies split their Aa/AA and A/A categories in three additional subdivision each (e.g. Moody’s AA includes Aa1, Aa2, Aa3). Aaa/AAA as well as lowest categories are not subdivided.

In classifying bond issuers and obligations, the rating companies employ several of the mechanisms used in equity valuation, however, their approach centers on a longer time horizon, rather than in the short term perspective of equity holders, as agencies look for the interests of bond tenants and other creditors.

Most agencies are cryptic about the method utilized in their rating process, but they seem to monitor and calculate performance indicator in common areas including: Industry risk, Competitive position, management quality, financial leverage and flexibility, profitability, liquidity, among others. For instance, in evaluating a company’s financial robustness Fitch calculates a set of financial ratios, tracking them across a relevant period of time and combines this information with qualitative insights.

Typically, the data produced by rating agencies includes tables with the default experience of specific ratings during 20-year periods. An example of such products is reproduced in Table 2.2. This table presents the probability of a security defaulting during a certain year. For instance, a bond with a rating of A will have 0.520% probability of defaulting by the end of the fourth year.

It is noteworthy to mention that for investment level bonds, the probability of default increases as time elapses; while for speculative bonds this probability decreases. The reason behind this is that high quality issuers are initially considered creditworthy, but as time increases the likelihood of running into financial problems also grows. For junk bonds, on other hand, the situation is the
opposite because as time passes, the longer the company survives the higher the probability its financial condition will recover.

Table 2.1.

<table>
<thead>
<tr>
<th>Rating</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Investment Grade Ratings</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Rating</strong></td>
<td><strong>Interpretation</strong></td>
</tr>
<tr>
<td>Aaa/AAA</td>
<td>Highest quality; Extremely strong capacity to meet financial commitments (lowest credit risk); practically no probability to be affected by foreseeable events</td>
</tr>
<tr>
<td>Aa/AA</td>
<td>Very high quality; very strong capacity to meet financial commitments (very low credit risk); capacity for repayment not significantly vulnerable to foreseeable events</td>
</tr>
<tr>
<td>A/A</td>
<td>Upper-medium quality; strong capacity to meet financial obligations (low credit risk); more likely to be affected by changes in economic conditions</td>
</tr>
<tr>
<td>Baa/BBB</td>
<td>Medium grade quality; adequate capacity to meet financial obligations (moderate credit risk); negative changes in environment may affect payment capacity</td>
</tr>
<tr>
<td><strong>Below Investment Rating (Junk Bonds)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Rating</strong></td>
<td><strong>Interpretation</strong></td>
</tr>
<tr>
<td>Ba/BB</td>
<td>Considered speculative; subject to substantial credit risk</td>
</tr>
<tr>
<td>B/B</td>
<td>Considered very speculative grade; subject to high credit risk</td>
</tr>
<tr>
<td>Caa/CCC</td>
<td>Considered highly speculative (of poor standing); subject to very high credit risk</td>
</tr>
<tr>
<td>Ca/CC</td>
<td>Wildly speculative; likely to be, or very near, to be in default with some prospect of recovery of principal and interest</td>
</tr>
<tr>
<td>C/C/D</td>
<td>In default or bankruptcy</td>
</tr>
</tbody>
</table>

Hazard rates.

As we mentioned above, the probabilities included in Table 2.2 are historical probabilities of default at the end of a certain year. This information can be used to calculate the probabilities of defaulting during a particular year, which is known as unconditional probability of default. For instance, let’s estimate the probability that a Ba-rated bond will default during the fifth year as

\[ 10.397 - 8.123 = 2.274\% \]

this is the probability of defaulting in year five independent of what happened before. However, we may be interested in calculating the probability of the firm defaulting on the fifth year given that it survived on the previous periods. To achieve this we need first to obtain the probability of survival until end of year four, which is 100-8.123= 91.877%. Dividing the unconditional probability obtained above by this number we obtain

\[ 2.274\%/91.877\% = 2.475\% \]

which is the conditional default probability of defaulting on the fifth year. Such conditional probabilities of default are known as default intensities or hazard rates.

Hazard rate models are applied to estimate the default probability of firms with liquid debt securities, where an excess yield or spread of corporate bond yield is calculated over the return of equivalent risk-free bond. In practice, the reference risk-free used in quoting corporate bond
spreads is the yield on Treasury bonds with similar terms. This spread can be understood as a risk premium required by bondholders for the danger of default that they are incurring\textsuperscript{3}; therefore, using non-arbitrage arguments a risk neutral default rate can be calculated.

**Market models**

When public information exists about the company under analysis, and assuming the existence of efficient markets, a firm’s stock value, its variance, as well as its liabilities can be used to determine an estimate of the probability of default. As explained in a publication by Moody’s Global Credit Research\textsuperscript{4}, credit risk models relying only on historical financial figure “may present an incomplete or distorted picture of the company’s true financial condition” given financial statements’ goal to comply with the inherent conservatism of accounting principles or simply because accounting fundamental principle is to provide pictures about the past of a firm and not its future.

Nevertheless, as any other approach, public firm risk models have pros and cons. As we mentioned above and will explore in detail in our discussion of structural models, these type of models rely importantly on strong assumptions on market efficiency and the idea that equity prices incorporate relatively complete information about a firm’s performance quality and creditworthiness. Therefore, the model’s ultimate usefulness will depend on how well such assumptions reflect the true and whole dynamics of a company. Moreover, by definition, this

\textsuperscript{3} Some authors include a liquidity component in explaining the difference between corporate bonds and equivalent Treasuries (See Hull 2012, pg. 529)

approach has application not only on firms with available public information but only on those which possess sufficiently liquid stock prices\(^5\).

Recognizing that the information incorporated in the markets can be highly useful and that a scrupulous analysis of a firm’s financial statements remains a crucial element in evaluating the credit quality of corporate obligors, some firms have developed hybrid approaches combining the strength of structural public firm models and the financial statements’ insight of reduced-form statistical models.

A clear example of this kind of hybrids is Moody’s Public Firm Risk Model which puts together a variation of Merton’s contingent claims approach to credit risk with a non-linear statistical model regressing on company financial information, agency ratings, and other macroeconomic and capital markets data. The result is a one-year Estimated Default Probability (EDP), used as an early warning to monitor changes in corporate credit quality.

**Exposure Models**

As explained in the sections above, three main components affect the credit risk of a financial obligation\(^6\): 1) the probability of default (PD); 2) the loss given default (LGD); and 3) the exposure at default (EAD). While much work has been dedicated to PD modeling, LGD and EAD still lag behind in terms of both theoretical and practical insight.

This lag is mainly the consequence of a lack of reliable historical data which has to do with current operational practices and absence of systems to track information about events occurring

\(^5\) According to Moody’s, in United States, this subset of firms refers to less than 10,000 public companies.

\(^6\) Expected Loss (EL) = Probability of Default (PD) x Loss Given Default (LGD) x Exposure at Default (EAD)
after default. As explained by Stein (2009), there are number of reasons why LGD modelling, in particular, has fallen behind: 1) the definition of LGD changes across applications, institutions, and markets complicating the construction of standardized data bases and common approaches; 2) a historical lack of theoretical models, as compared to the PD prolific area, around which to build equivalent theoretical constructs, 3) a high sensitivity of LGD calculation to the assumptions used to estimate it, and 4) the predictive power of any LGD model depends on hard to quantify variables (e.g. legal resolutions in bankruptcy processes). Despite all these obstacles, accurate predictions of LGD and EAD are essential elements that should be calculated to achieve a truly comprehensive risk management strategy. Renewed attention in LGD/EAD data collection and modeling started in 1999 when specific mandatory Basel requirements were imposed upon financial institutions to become compliant with advanced internal rating bands and have been improving ever since.

In the following lines, we will present the main elements of exposure models. EAD models calculate how much is at risk in a given facility (loan exposure) conditional on a default event. These figures are particularly important for creditors granting credit lines, although the models are also useful for other types of securities including swaps, caps and floors. In the case of lines of credits, EAD is divided into drawn and undrawn commitments. While the drawn component is typically known, the undrawn part needs to be estimated to arrive at an accurate estimation of EAD. Typically, EAD calculations present the following structure:

\[ E[EAD] = E[EAD]_{\text{cash}} + E[EAD]_{\text{undrawn}} = [DDF_{\text{cash}} \times E[X] \times \text{Cash}] + [DDF_{\text{undrawn}} \times CEEF \times \text{Undrawn balance}] \]
where; DDF = draw-down factor for cash and contingent portions respectively; E[X] = average use of available cash; CEEF = cash equivalent exposure factor for contingent part (i.e. conversion factor of undrawn liability into a cash exposure).

2.2. Techniques used in the measurement of credit risk.

Within the default prediction discipline, models can be categorized based on three main dimensions: the area of applications (which we explored in section 2.2.), the products to which they are applied (in the present work our focus is low-income consumer lending), and on the type of techniques utilized. In the following lines, we will provide an overview of the main techniques used in the industry today.

As explained by Hand and Henley (1997), the quantitative applications used to discriminate between good and bad borrowers were dominated in its beginnings by basic statistical approaches, particularly by the technique known as Multiple Discriminant Analysis (MDA). But, given the restrictive assumptions behind MDA, the techniques evolved to more sophisticated methods including: econometric models, advanced mathematics and optimization procedures, and neural network approaches. No matter which technique is employed, the reader should remember that the capacities of any default prediction model depends on the quality of the information used (garbage in-garbage out).

Among the more often used techniques we have:

1. **Econometric techniques.** Unlike structural models, econometric models do not rely necessarily on a causal relationship to explain an obligor’s probability of default; however, it
is important to mention that the selection of variables follows some type of economic intuition or underlying economic logic. All the econometric models used in the industry (whether they be Logit or Probit analysis, multiple regression, or linear and multiple DAs) have the probability of default (and in other cases the default premium) as the dependent variable whose variation is explained by a set of independent factors. Most commonly, the explanatory variables can include (depending on the unit under analysis): financial ratios, other internal performance indicators (in the case of consumer lending credit reports and scorings), as well as external variables such as macroeconomic factors, industry indicators and/or agency ratings. In the case of Hazard rate/Duration models (also known as survival analysis), the econometric techniques are used to calculate the time to failure, death, or certain response to a specific event. These two techniques (together with CART, Random Forests and Optimization techniques) will constitute the methodological core of our research, thus we will dedicate Section 2.4 for a more detailed discussion on these methods.

2. Optimization Models. Among the main methods used in credit scoring we have: linear, quadratic programming, integer, multiple criteria, and dynamic programming. As explained by Caouette, the objective of these mathematical programming techniques in the field is to find the optimum weights for obligor and credit attributes that maximize profits for creditors, while minimizing errors correctly classifying customers. The works of Hand (1981) and Kolesar and Showers (1985) are among the first models that were applied, and establish its applicability, in the credit scoring industry. There have been also a number of authors that have expanded this methodology to predict the delay in credit payments.

3. Neural Networks (NN). Categorized within the nonparametric group of credit scoring methods, this technique tries to simulate the human thought process and methods of learning
by emulating an interconnection of basic units of decision (nodes) that can compute values from inputs by feeding information through the network to reach to a credit scoring decision. Neural network algorithms consist on a group of inputs that are mathematically transformed through a transfer function to generate a result. As input nodes we have variables related to the credit particulars (e.g. borrower attributes, environment information, etc.), this information is then processed in the system producing an exit node defined as the probability of default. Among other important elements in NN, we have: type of model, transfer function, number of layers, amount of processing components in each layer, and learning rule parameters to modify weights in the process. During the training process, the weights are adapted to minimize differences between the desired and actual output (i.e. right prediction of default and generated prediction). In numerous works, including Richeson, Zimmerman and Barnet (1996), and Jensen (1996), the discrimination accuracy of NN model ranges from 76 to 82%. Despite its accuracy, it is noteworthy to mention that the credit scoring process through the use of neural networks is somewhat cryptic, as the inner learning process works as a “black box” and it is somewhat difficult to explain the predictions in terms of inputs. This could imply an obstacle from a regulatory point of view in terms of the transparency and tractability of results. This is one of the reasons why NN has been more applied in the backend (post-approval credit evaluation) rather than in the front-side (credit granting) of the field. In 1994, Ripley described some additional applications of neural networks for fraud detection and managerial decisions. The application in this area has had broader acceptance given that it is more important to accurately detect fraud than explaining in detail the variables that were used in arriving the decision. Nevertheless, computational and
methodological improvements make NN a promising technique both in current and new areas in the credit industry as a whole.

4. Decision Trees (DT). Classifications and Regression Decision Trees, Random Forests and other decision trees techniques (e.g. ID3, C4.5, and C5) are part of the operations research tool-box. One of the main advantages of this technique is that it is not restricted to statistic assumptions on the form of the distributions or functional forms. In the decision tree process, the borrower’s attributes are split-off consecutively in branches from most to less important classifying the population according to these partitions. For instance, in a first split, a borrower’s population might be divided in two main groups based on whether they own a house or rent. Subsequently, the rent sub-group can be further subdivided according to their age, or to its declared income. This process continuous until the population is classified into mutually exclusive buckets. The credit decision then can be taken based on the probability of defaulting (target variable) in each of such buckets. Although decision trees possess advantages in terms of interpretation over other techniques (for its visual relationship between attributes and target variable), there might be some drawbacks if the splits are too few to be used by credit managers not only as a prediction device but as a decision guide in real operations. However, DTs continues to be also a promising tool in credit scoring, including the use of more sophisticated approaches like Random Forests (RFs) designed to improve the prediction accuracy of CART. Random Forests works by building a large number of CART trees (although this makes the model less interpretable), where each tree issues a “vote” on the predicted outcome, picking the outcome that receives the majority of votes. The improvements in accuracy of RFs over regular decision trees are considerable, and there is plenty of opportunity to explore its applications in credit scoring models.
5. **Rule-based models.** In this type of method the objective is to simulate the decision making process followed by a successful credit analyst. This methodology allows to build a structured decision process, where best practices, experience and business policies are standardized and translated into a logical set of business rules. The elicitation of a *champion* analyst’s decision process enables the dissemination of this expertise across the organization. Also known as expert models, this technique consists of: 1) a group of decision algorithms, which is applied to a knowledge base including relevant credit data (e.g. financial ratios of credit applicant), and 2) a structured inquiry procedure used by regular analysts to get information of a specific client.

6. **Hybrid models.** Combining different approaches (including simulation, statistical estimation, and other mathematical computation techniques) these models try to take advantage of the strengths of various methods to reach a more accurate credit scoring prediction. For instance, Chuang and Huang (2011) develop a two-stage model using NN for an initial categorization of borrowers into good and defaulting categories, and then switching to a case-based reasoning (CBR) classification technique to reduce the ratio of rejected good applicants (Type I error) in the original step. An additional example of hybrid systems is Moody’s KMV model, which is based on a structural approach and estimation procedure to calculate *Expected Default Frequency* (EDF). To achieve this, KMV transforms a default probability produced by Merton’s model into a real-world default probability (EDF), calculating also the Distance to Default and using an empirically derived default point (defined as $F = \text{Short Term Debt} + \frac{1}{2} \text{Long Term Debt}$).
2.2.1. Econometric Models, Classification and Regression Trees, and Random Forests

Econometric Models: Introduction and Historical Review

The literature about PD models based on econometric techniques is vast and substantial. For over forty years now, numerous authors have examined and refined approaches to predict default risk using quantitative measures of performance (e.g. liquidity, profitability or leverage) to identify statistical relationships with credit events. However, this approach is relatively new. Without any doubt, the primary works in this area were those by Beaver (1967) and Altman (1968), who regressed univariate and multivariate models on financial ratios to predict company’s failures.

In his original work, Beaver run a binary classification test to determine the error rates a potential lender would experiment if the classified borrowing firms as failed/non-failed on the basis of financial ratios. To do this, the author used a matched sample consisting of a total of 158 firms (79 healthy and 79 failed) and analyzed fourteen financial ratios on this companies. Beaver’s original models were univariate, analyzing single historical financial ratios five years before failure and looking to discriminate between non-failed and failed companies. In this model, Beaver identified that an individual financial ratio (Cash Flows/Total Debt) had the statistical power to differentiate between healthy and defaulting firms, however this univariate approach presented inconsistency problems.

A year after Beaver’s revolutionary work, Edward Altman conducted a multiple discriminant analysis approach (MDA) that solved the univariate model’s inconsistency problem and that assessed a more integral financial profile of companies. To this day, this article is considered the
cornerstone of statistical failure prediction models. In his study, Altman examined also a matched sample consisting of 66 manufacturing companies (again 33 healthy and 33 defaulted) and exploring 22 potentially explanatory financial ratios. In his final model, the author ended up with five ratios providing the highest statistical prediction of corporate bankruptcy. His innovation consisted in utilizing a set of multiple factors to explain default, rather than only one, as a combined set of forces is what contributes to the failure of a company. The variables used in this study were classified into five main ratios groups: leverage, solvency, profitability, activity, and liquidity.

For many years after this study, MDA remained as the predominant technique used in failure prediction models, although two central MDA assumptions (independent variables follow multivariate normally distribution, and variance-covariance matrix are equal among healthy and defaulting groups) were regularly violated. Furthermore, MDA’s coefficients presents problems of interpretation. MDA’s coefficients cannot be interpreted as slopes of a regression equation, thus they do not denote the marginal effect of a factor over the dependent variable (relative importance). As we will see below, in 1980, J.A. Ohlson proposed the use of a Logit model bypassing MDA’s restrictive assumptions and the need of proportional samples; nevertheless, the Altman’s model, known as the Z-score still retains its honor place as the prototypical default prediction model. A more detailed discussion of discrete-choice models will be conducted in the next section.

To this day, the Z-score model has acquired a benchmark status for its parsimony and predictive power. In numerous comparison exercises, Altman’s approach consistently beats or at least is at a draw when contrasted against other models. As will be discussed in coming sections, the Z-score’ ROC and AUC are hardly surpassed. In the last years, the Z-score has also been updated,
improved and calibrated. Among the recent revisions we can count: log transformations of several raw factors, new explanatory variables on size and age (scaled by total stock market size), and recalculation of model using logistic regressions rather than MDA.

**Econometric Models in Credit Risk Modeling**

Econometric models possess a preponderant position in the credit risk modeling practice. Within the industry two techniques are dominant: *duration models* and *discrete-choice models* (including Logistic and Probit Regressions). Although the use of the latter is more widespread in the industry, the former is acquiring followers given its more efficient use of scarce default observations and more flexible analytic framework over various horizons with the same model. In the coming lines we will present general description of these two models (focusing on Logistic regressions which will be the main econometric method utilized in this research).

**Logistic Regression**

Like linear models, Logistic regressions tries to identify a statistical relation between a series of explanatory variables and an outcome (dependent variable), the main difference is that Logit models is applied in situations where the dependent variable is categorical. In the case of credit risk modeling this variable often is defined as a binary variable where 1 represents the default event and zero is a non-defaulter. As in the case of linear models, in Logistic regressions our independent variables can be numerical, categorical or a combination of both. Whereas in linear models the objective is to predict the value of a continuous dependent $Y$, in Logit our aim is to forecast to which class will a new observation will belong (or just to classify it into some defined category). Logistic regressions consists of two main stages: 1) obtain estimates of the probabilities of belonging in a specific category, and 2) determine a cutoff or threshold on these
probabilities to classify each case in one of the categories (which will be discussed in more detail below in our Confusion Matrices section).

The Logistic regression derives its name from the use of a function of the dependent variable rather than the objective variable itself. As we will see below, this function is called the Logit which will be modeled as a linear function of predictors. Once the Logit has been forecasted, we can then map it back to a probability form.

To reach this Logit function, first we look at the probability $p$ of belonging to class 1 (as opposed to class 0). Unlike a normal numeric dependent variable $Y$, here $p$ can only take values in the $[0, 1]$ range. This creates a problem because if we define $p$ as a linear function of a group of independent variables in the form:

$$ p = B_0 + B_1 x_1 + B_2 x_2 + \ldots + B_q x_q, $$

We are, by no means, assured that the set of independents will lead us to a values between zero and one. The solution hence is to utilize a nonlinear function of this independent variables, as follows:

$$ p = \frac{1}{1+e^{-(B_0+B_1 x_1+B_2 x_2+\ldots+B_q x_q)}} $$

This is called Logistic response function and unlike the base form, this structure will always guarantee that the left side of the equation (i.e. the probability) will always be in our desired $[0, 1]$ interval.

Another important and related concept is known as the Odds, which is defined as the ratio of the probability is defined as the ratio of the probability of belonging to class 1 relative to the probability of belonging to the zero category. Numerically, the expression would be as follows:
Odds = \frac{p}{1 - p}

This metric is useful and popular because rather than talking about the probability of an event, is more intuitive to think in terms of the odds that an event will happen. For instance, if the probability of defaulting is 50% the odds of default will be 1 (=50%/50%). Thus, given the odds of an event, we can always calculate its corresponding probability. This relation is expressed mathematically as:

\[ p = \frac{odds}{1 + odds}. \]

Based on these past equations, we can re-formulate the relationship between the odds and independent variables as:

\[ Odds = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}. \]

This equation defines a proportional relationship between the independents and the odds, which can be interpreted in terms of percentages. That is, a unit movement in the independent variable \( x_j \) would be associated with a \( \beta_j \times 100\% \) increase in the odds. Applying logs to our past expression, we obtain the typical form of the Logistic regression:

\[ \text{Logit} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k. \]

; where Logit = log (odds), taking values from \(-\infty \) to \(\infty\).
This final form of the relation between the response and the independents uses Logit as our dependent variable, which can be model as a linear function. This will be the basic formulation we will employ in our econometric models.

**Duration Models**

According to Bohn and Stein (2009), one disadvantage of discrete-choice models as Logistic regression, is that it does not allow us to deal with "censored" data. As described by the authors, if default is described as an event occurring \( y \) years after a certain date, then any observation within that period will not have a complete period of time to express its failing or non-failing behavior. Duration models, on other hand, try to measure the length of survival (or time to death) instead than the occurrence of an event or not. In this case, rather than having the default event as our dependent variable our objective is to predict the time to default. There are a number of parametric (e.g. Exponential, Weibull, Log-logistic) and non-parametric versions of the baseline hazard function. Among the more used versions we have *Cox parametric proportional hazard model*, formally defined as:

\[
\lambda(t, v) = \lambda(t) \cdot e^{v\beta}
\]

; where \( \lambda(t) = \) baseline hazard rate.

It is noteworthy to mention than in Cox formulation the covariates act proportionally across the entire baseline hazard model. Moreover, this model exhibits duration dependence, i.e. the probability of default will depend on the period of time that the borrower has been at risk but has not defaulted.
Classification and Regression Trees (CART), and Random Forests (RF)

In the present section we will describe two additional methods that will be utilized to estimate the probability of default: Classification and Regression Trees, and Random Forests. It is important to mention, that given the nascent stage in which low-income consumer lending is, this kind of techniques have not been fully employed having Logistic regressions as the dominant method in the industry. As we will mention in detail in Chapter 3, a combined use of these three methods can provide credit managers with finer details that otherwise would have not been observable through the use of a single method alone.

Classification and regression Trees (CART) are among the most transparent and interpretable statistical methods. CARTs are basically based on a recursive partitioning process where observations in a sample are assigned into subgroups by creating splits on the independent variables, these splits then create logical rules that are easily comprehensible and can guide the modeler in understanding and applying the classification process.

There are two main concepts we will focus in our discussion about CARTs: 1) the concept of recursive partitioning behind the building of trees, and 2) the concept of pruning.

Let us discussed the idea behind growing a tree: recursive partitioning. In a classification problem, we want to forecast a class or response $Y$ from a group of predictors $(x_1 + x_2 + ... + x_q)$. To do this we develop a binary tree where at each internal split we apply a test to one of the predictors, based on the outcome of that test we go either to the right or left sub-branch of the tree. With the time, we reach to a lead node where we generate a prediction, which averages or
aggregates all the training data points that reached that leaf. One of the advantages of trees is that they do not assume a linear relationship among the dependent and independent variables. Many data sets have intricate interactions among their variables that usually follow nonlinear dynamics, thus generating a global model can be very difficult. Some solutions are to fit on-parametric models locally and then aggregating them, but the results tend to be very difficult to interpret. As we have mention, the alternative, thus, consists in partitioning the space into smaller more manageable sections. This division is then applied recursively, that is, operating on the results of the prior partitions, until we reach a point where have divided he original space in into \( n \) sub-areas where each sub-area is as pure as possible (i.e. it contains as much points as possible belonging to just one class). Once the areas are so tame we can then fit simple models and predict the dependent.

The second idea behind CARTs is the concept of pruning the tree. It is noteworthy to mention that as a tree grows it is more likely to create overfitting issues, hence, we will have be some “weak” branches hardly contributing to error reduction. This weak branches need to be eliminated or pruned. Formally defined, pruning involved in sequentially selecting a node and reclassifying it as a leaf (i.e. cutting all the branches below it). The trade-off then is one between misclassification error and tree complexity. This idea is more formally conveyed in the following expression:

\[
Cost \ of \ complexity = Error(\tau) + \alpha \times P(\tau)
\]

; where \( Error(\tau) = \) proportion of observations that are misclassified in the development set by the tree \( \tau \); \( \alpha = \) is the size “penalty” coefficient.
On other hand, closely related to CART models, Random Forest (RF) is another type of recursive partitioning method involving a group of classification trees (i.e. ensemble learning method) that are computed on random subsets of a sample utilizing a group of random predictors for each split in each CART. The results of Random Forests have been demonstrated to generate better and more accurate predictions than the results of standalone CARTs. The main idea is to create a large number of decision trees (therefore the name forest) and then having each tree issuing a “vote” that then is combined to obtain a sort of consensual prediction. Among the most important advantages of RFs are: 1) there is no need for pruning trees; 2) reduced sensitivity to outliers in the training data, 3) variable importance and accuracy are produced automatically.

2.3. Statistical Model Validation Techniques and Evaluation Criteria

In the following section, we will explore the main validation components used to ensure highest usefulness in model development. We will focus on three core dimensions: power, robustness, and calibration. Afterwards, we will present the principal measures and statistics used to describe how powerful and well calibrated a model is. Finally, we will discuss on measures to evaluate and correct PD model overfitting.

Validation Dimensions

Without question, the first criteria to evaluate the quality of a credit scoring model is its power to accurately discriminate between failing and non-failing borrowers. Nevertheless, while a powerful model by itself might be enough for “down and dirty” analysis, strategic portfolio management requires attention in other dimensions. Decision support systems can be qualified
across a broad variety of criteria ranging from model’s accuracy, ease with which results can be interpreted, or its ability to be adapted and modified to changing needs. The work of the credit modeler is thus to select the best combination of attributes that ensure a model not only with the power to discriminate between good and bad borrowers but also with a well calibrated, dynamic and parsimonious structure that facilitates its use in daily credit scoring operations.

Stein (2009), and Kumra, Stein and Assersohn (2006) highlight the following dimension to assess a model’s performance:

a) Power: In terms of credit scoring, power is defined as a model’s capacity to accurately differentiate between defaulting and non-defaulting borrowers. Model power, therefore, must be our departing point and most fundamental objective. As we will discuss in the calibration section, having powerful predictions per se is not enough for many critical credit applications, in these cases accuracy and precision is also fundamental. But despite this, we must be aware that powerful models can most of the time be improved through adequate calibration, while a perfectly calibrated model will be almost useless if its prediction power is poor. Power hence must be evaluated first.

b) Robustness: As we mentioned above, it is not enough that a model predicts well but also that its performance is consistent over time. This is that its parameters are sufficiently stable so that its prediction and relative importance of variables does not change drastically from sample to sample and time to time. Model robustness is defined as a model’s sensitivity (both of parameters and predictions) to changes in the data.

c) Calibration A well calibrated PD model is one where its outputs are unbiased predictions of the real probabilities of default (i.e. that its conditional predictions of default accurately forecast the actual default rates). As we will see when discussing detailed calibration
techniques, a valid concern about mechanical credit rating systems is that they can be volatile, i.e. a prediction might change when modifications in the underlying data happened but no fundamental change has occurred. This results in noise. Our ultimate objective in calibration, thus, is to adjust the model’s estimated PDs to correspond to actual outcomes as tightly as possible. Calibration in PD models is required, because default information does not follow regular distributions and many statistical assumptions are not met. Nevertheless, the ordering of PDs from the model is still strongly correlated with the actual ranking of real PD observations. This strong relation implies that a transformation can be applied to our model’s mapping to better resemble true probabilities. This is the goal of calibration. Finally, as mentioned in the power discussion above, if a model’s primary objective is just to rank the risk of borrowers, poor calibration will not represent a huge issue (as long as the model is powerful enough); but, when accurate predictions are a central part of decisions, such as in security valuation or capital allocation, both power and calibration will have to be considered. As a result in this type of applications modelers habitually spend a significant amount of time adjusting models to true probabilities.

d) **Transparency and Diagnosability:** These two elements refer to the extent to which users can review the model’s structure, parameters, and outputs and understand in a general way how that model works. Transparency is related to how easy interactions between inputs and outcomes can be traced and deduced, while diagnosability refers to the capacity to dig into the model structure and derive conclusions about its dynamics in an efficient and expedite way. As mentioned above, credit modelers are constantly faced with a choice between usability and complexity. On one hand, a complex model can provide higher discrimination power, but this accuracy can be at the cost of interpretability and speed. On other hand,
having a refined complex prediction might be desirable if speed is not the critical variable and if and only if diagnosability is not sacrificed. The final decision depends on the specifics of each situation, although the consensus among most practitioners is toward parsimonious specifications.

Given the importance of accurate PD estimates, in the following lines we will elaborate on some of the most important tools used to validate a model’s *power, robustness, and calibration*.

**Model Power tools: Confusion Matrices**

As we mentioned above, in credit classification models the ability to discriminate between good and bad loans is critical, this mean how well a model classifies each borrower in the “correct” category. Formally speaking, the most powerful model will be the one with more true non-defaulting and less false negatives in its “good” category, and with more true defaulting loans and less false positives in its “bad” group. This type of analysis is known as *confusion, classification* or *contingency matrices*.

According to the definition provided by Shmueli (2010), a contingency matrix is a tabular representation of categorical data displaying the multivariate frequency distribution of a group of variables. In the case of PD models, the confusion matrix is used to compare the number of predicted failing (non-failing) loans to the “real” number of failing (non-failing) credits observed. A summary with the general structure of a confusion matrix is presented in Table 2.3.

In this table, a *True Positive* would be any predicted default that occurred in reality, while a *True Negative* would be any predicted non-failing credit that actually did not default. A *False
Negative represents any predicted credit that was classified as non-default by the model, when in reality was an actual default, and a False Positive is any credit that the model predicted was a defaulting loan when in reality was a non-default.

In this table, we also present two measures that will be fundamental for other of the power valuation tools that we will explore above (i.e. ROC or power curves), this measures are: 1) Sensitivity: defined as the percentage of True Positives relative to the total number of actual defaults (TP + FN), and 2) Specificity: defined as the True Negatives predicted by the model divided by the actual number of non-defaults observed in reality (TN+FP). This two measures will be used to calculate the True positive rate and False Positive rate that are the main components of a Receiver Operator Characteristic (ROC) curve.

Table 2.3. Typical confusion matrix structure

<table>
<thead>
<tr>
<th>Actual Non-Defaults</th>
<th>Predicted Non-Defaults = 0</th>
<th>Predicted Defaults = 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Negatives (TN)</td>
<td>False Positives (FP)</td>
<td></td>
</tr>
<tr>
<td>False Negatives (FN)</td>
<td>True Positives (TP)</td>
<td></td>
</tr>
</tbody>
</table>

*Source: own elaboration

TRUE POSITIVE RATE (Sensitivity) = TP / (TP + FN)
FALSE POSITIVE RATE (1 - Specificity) = 1 - [TN / (TN +FP)]

Model Power tools: Receiver Operator Characteristic curve (ROC) and Power Statistics

From our discussion on confusion matrices above, it is noteworthy to mention that for default models generating continuous outputs, such as PDs, a different threshold (or cutoff) value will change the number of observation in each types of errors.

To exemplify this problem, in Table 2.4, we present different thresholds and how they modify the number of observations within each category (See how the Sensitivity and Specificity
measures change with different cutoffs). This presents a challenge with the relatively arbitrary selection of thresholds, thus which threshold should we select? As will be explained below, among our power valuation toolbox we have an instrument that allows us to that captures all thresholds simultaneously: the ROC curve.

Table 2.4. Example of Threshold changes in contingency tables

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Predicted Non-Defaults = 0</th>
<th>Predicted Defaults = 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Non-Defaults</td>
<td>54</td>
<td>20</td>
</tr>
<tr>
<td>Actual Defaults</td>
<td>9</td>
<td>16</td>
</tr>
</tbody>
</table>

Sensitivity = 16 / 25 = 0.64
Specificity = 54 / 74 = 0.73

Threshold = 0.5

<table>
<thead>
<tr>
<th>Predicted Non-Defaults = 0</th>
<th>Predicted Defaults = 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Non-Defaults</td>
<td>70</td>
</tr>
<tr>
<td>Actual Defaults</td>
<td>15</td>
</tr>
</tbody>
</table>

Sensitivity = 10 / 25 = 0.4
Specificity = 70 / 74 = 0.95

Threshold = 0.7

<table>
<thead>
<tr>
<th>Predicted Non-Defaults = 0</th>
<th>Predicted Defaults = 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Non-Defaults</td>
<td>73</td>
</tr>
<tr>
<td>Actual Defaults</td>
<td>17</td>
</tr>
</tbody>
</table>

Sensitivity = 8 / 25 = 0.32
Specificity = 73 / 74 = 0.99

*Source: adapted from Bertimas, D.; Analytics Edge course materials (2014)
Receiver or Relative Operator Characteristic (ROC) curve is a graphical plot that shows the performance of a binary classifier system (model) as its discrimination cutoff varies. In the case of PD models, ROCs are built by scoring all the credits and then sorting them from worst to best on the horizontal axis (FPR axis), and then drawing the fraction of defaults excluded at each level on the TPR axis. Therefore the ROC curve is formed by plotting the True Positive Rate (See Table 2.3) against the False Positive Rate at various threshold settings. ROC is simply the Sensitivity as a function of (1 – Specificity).

Figure 2.1. Receiver Operating Characteristic (ROC) curve

Figure 2.1 presents an example of how a typical ROC curve looks like. The best possible model thus would provide a point in the coordinate (0, 1) indicating a perfect sensitivity (no false negatives) and zero specificity (no false positives), while a random guess would give an
observation along a 45 diagonal in the 1x1 square. This diagonal known as no discrimination line indicates that points above it represent good discrimination results while points below it represent results worst that random (it is curious to mention that the result of a consistently “wrong” predictor could just be inverted to obtain a mirror good predictor). It is important to mention that when a model’s ROC consistently lies above a second model’s ROC curve, then the strictly dominating curve will provide lower errors at any threshold selected than the model with the ROC below it. On other hand, if we have two models whose ROCs flip each other (e.g. Model x’s ROC is above Model y’s ROC below a certain point and then Model y’s ROC becomes higher than Model x’s ROC from that point on), then we would use the model with the higher ROC in the range of credits we are interested in (i.e. Model x if we want to identify defaulters in low quality borrowers, or Model y if we are interested in detecting classifying credits in higher quality applicants).

Sobehart, Keenan, and Stein (2000) developed an analogous measure called Cumulative Accuracy Profile (CAP). While the ROC curve plots the proportion of defaults below a credit score x versus the proportion of non-defaulters below that same score x, the CAP plots the proportion of defaulters with a score below x vis-à-vis the proportion of all the firms in the sample with a score below x. ROC therefore tells us how much of a non-defaulters have to be excluded to avoid a certain percentage of defaulters, while CAP tells us how much of the total portfolio has to be excluded to avoid a certain proportion of defaulters.

**Model Power tools: Area Under the Curve (AUC) and Accuracy Rate (AR)**

Another useful tool to evaluate the power of a model is represented by the area under the ROC curve as a proportion of the unit square as shown in Figure 2.1. This measure is known simply as
AUC and tells us that if we are given two observations, one defaulter and one no defaulter, (and assuming AUC=0.75) our model would be able to correctly detect its status in AUC/unit square (= 3/4) of the times. This is way better than a random classification. In the case of ROC, and as mentioned above regarding the non-discrimination diagonal, an AUC of one would indicate a perfect model while AUC equal to 0.5 imply random discrimination. Figure 2.2 presents these two cases graphically (compared to a model’s AUC of 0.775).

Following the definition of Sobehart, Keenan, and Stein (2000), a similar metric called the Accuracy Ratio (AR) can be calculated for the CAP where its values range from one (perfect prediction) to zero (random classification). As in the case of AUC, AR captures a model’s predictive accuracy for each type of error by comparing how close a model’s CAP is to the data set actual (ideal) CAP. To achieve this the area between our model’s CAP and a random classification CAP is divided by the area between the ideal CAP and the random CAP (maximum area possible).

Figure 2.2. Maximum/Minimum AUC values

---

7 i.e. AUC tells us the probability that a model will rank a randomly selected positive observation higher than a randomly chosen negative (and under the assumption that positive ranks higher than negative).

8 Mathematically, the Accuracy rate is expressed as: $AR = \frac{\int_{b}^{1}y(x)dx - 1}{1-f} = \frac{1 - 2\int_{f}^{1}z(x)dx}{f}$; where: $y(x)$ and $z(x) = are Type I and Type II CAP respectively; $f = is the proportion of failing loans to total number of observations (d/[n+d])$
An additional performance metric is the Conditional Information Entropy Ratio, which captures in a single statistic the number of failing credits included in the distribution of model scores (i.e. Information entropy). The CIER is based on the Information entropy (IE) measure which is a measure of the overall uncertainty represented by a specific probability distribution. The first step is to calculate IE without controlling for information we may have about credit quality (Ho), this entropy will reflect the likelihood of an event given by the probability of default (knowledge common to all models). The second step is to calculate the conditional information entropy (H1[s, δ])10, which takes into account the probability that a borrower will default given that his risk score is R. The CIER, thus is defined as:

\[ CIER(R) = \frac{H_o - H_1(s, \delta)}{H_o} \]

9 Mathematically, IE is defined as:
\[ H_o = p \log(p) + (1 - p) \log(1 - p) \]
where: p = probability that an obligor will default; and (1-p) = probability it will not default. For a detail exposition of the topic see Keenan and Sobehart (1999) [Keenan, S., and Sobehart, J.; Performance Measures for Credit Risk Models; Research Report; Moody’s Risk Management Services; 1999]

10 Formally, CIE is defined by:
\[ H_i(s, \delta) = \sum_{k=1}^{n} h(R_k) * p(R_k) \]
where: \( h(R_k) = -(p(A|R_k) * \log p(A|R_k) + p(B|R_k) * \log p(B|R_k)) \)
A = event where obligor defaults; B = event where obligor does not default; S = \{R_1, ..., R_n\} is a set of scores; and \( \delta= \) the size of the buckets in which the model output is divided.
Given that the CIER measures the reduction on uncertainty, a larger value represents a better model. Therefore, if the CIER is close to zero, then the model has almost no predictive power (i.e. the model gives no additional information about default occurrence that was not known previously). On the other hand, if CIER is close to 1.0, then we would have perfect prediction (i.e. no uncertainty would exist about the default prediction).

**Calibration techniques: calibration to PDs**

As discussed above, although model power is the primary objective of any credit scoring model, there is a number of credit applications where discriminating power alone is not enough. In credit valuation and portfolio management, for instance, precision and accuracy, not just power, are fundamental. In valuation decisions, poorly calibrated PDs can produce price estimates diverging importantly from actual observations, while in risk portfolio management, they can yield deceiving estimates of the size and distribution of losses. As we see, accuracy in this type of applications is essential as small differences can translate in important losses impacting directly the profitability of companies. However, calibration becomes important only after power has been achieved. As stated above, powerful models can be improved through adequate calibration (most of the time), but low power models will never improve their predictions if well calibrated. Power, hence, is always our starting point.

In default prediction models, calibration refers to adjusting a model’s predicted PDs to match actual default rates as closely as possible. As noted above, calibration is required because credit information does not follow the regular distributions required in econometric approaches or sample data is not fully representative of population, then PD predictions will tend to be biased.
(consistently lower or consistently higher). However, the model’s credit ordering will be still strongly related to the actual ordering of real probabilities, which implies that if we can transform the “rankings-probabilities” mapping (keeping intact the ordering) then the model’s output can be modified to better reflect the real probabilities. This use of transformations to address non-linearity problems and data noise was introduced by Falkenstein (2000) and ever since has been widely adopted in the credit modeling field.

Following Bohn and Stein (2009), this probability calibration involves a two-step process. In the first step, we map our model’s probabilities to empirical PDs using historical data. To achieve this, we order all credits by scores and create \( n \) buckets, where the lower quantile will contain the observations with highest probability of default and the \( n \)th will include the loans with less default probability. Then we estimate a transformation function mapping buckets to their historical default rates using nonparametric smoothing techniques.

In the second step, we adjust for differences between default rate in historical data and real rates (i.e. we adjust to reflect prior probabilities distribution). It is important to mention that this step is necessary only when our sample’s mean default rate differs importantly from the population’s mean default rate due, for instance, to missed or incomplete development data sets. In the case that we have complete data this stage would be redundant. Following the methodology used in Falkenstein (2000), a simple way to achieve this adjustment is by netting out the model’s rate and then multiplying it by the true baseline probability. For instance, if the sample default rate were 40bps and the populations’ correct rate were 200bps, our model’s PDs would have to be adjusted by a factor of 5. However, this simple approach can present a problem when sample mean PD rates are bigger than populations’ default rates. In this case, the adjusting factor would
be larger than the unit yielding adjusted PDs larger than 1 (which is impossible). This can be corrected whether imposing artificial limits on model’s probabilities or utilizing the following adjustment proposed by Elkan (2001):

\[
p_i^{adj} = \pi_{real} \frac{p_i - p_i \times \pi_s}{\pi_s - p_i \times \pi_s + p_i \times \pi_{real} - \pi_s \times \pi_{real}}
\]

where: 
- \( p_i^{adj} = \) final adjusted PD 
- \( p_i = \) original PD obtained from the model 
- \( \pi_{real} \) and \( \pi_s \) are real and sample probabilities of default correspondingly.

Calibration techniques: Likelihood metrics

So far we have discussed calibration using discrete buckets. In the following lines we will discuss the continuous scenario, i.e. when the size of the quantiles becomes infinitesimally small. As in our past discussion, our objective is to evaluate the match between the actual PDs and those predicted by our model. In this case, likelihood estimates will provide a measure to compare the accuracy of competing models, defining as the best option that model whose proposed PD mean is probabilistically more consistent with the mean PD observed in the population.

A numerical example will help us to present in more formal terms these ideas. Let us assume that defaults follow a binomial distribution with value of 1 if default is observed and zero otherwise.

---

11 In normal cases the adjustment function would be defined by: \( p_i^{adj} = (\pi_{real}/\pi_s) \times p_i \)

12 This can be achieved defining boundaries on the transformation of the model’s PD or selecting a functional form of the model that establishes an upper bounds to its outputs.
Following the traditional specification of binomial probability function (See Wackerly, Mendendall and Scheaffer (2002)) we know that the likelihood of observing \( y \) defaults out of \( n \) observations is given by:

\[
p(y) = \binom{n}{y} p^y (1 - p)^{n-y}
\]

where: \( p = \text{is the probability of default proposed by the model} \); \( y = \text{is the number of defaults observed in the population} \); \( n = \text{is the total number of observations in the population} \)

Moreover, let us assume that we are interested in evaluating the PD predictions of two models (Model A and Model B) to estimate the PD of a population of 200 credits with 10 observed defaults. Model A proposes a 4% average PD\(^{13}\), while Model B proposes a 10% average PD. Table 2.3 summarizes the likelihood calculations.

Table 2.3. Example of likelihood estimate for two models’ proposed average PDs

<table>
<thead>
<tr>
<th>Model’s proposed avg. PD</th>
<th>Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu_A = 4% )</td>
<td>( p(\mu_A = 4%) = \binom{200}{10} (4%)^{10} (96%)^{90} = 0.101 )</td>
</tr>
<tr>
<td>( \mu_B = 10% )</td>
<td>( p(\mu_B = 10%) = \binom{200}{10} (10%)^{10} (90%)^{90} = 0.005 )</td>
</tr>
</tbody>
</table>

\(^{13}\) Given that correlation is present in credit data (e.g. economic cycle effects), non-diversifiable skewness will appear in the default distribution (i.e. the median will always be smaller than the mean). This implies that comparisons using the mean default rate will typically overstate PDs. An alternative approach developed by Kurbat and Krobalev (2002) uses median number of defaults instead. To achieve this, the authors propose simulating the default distribution using a factor-based approach and calculating the median of this simulated distribution. This work is part of a family of procedures acknowledging the presence of correlation in default samples, due for instance for a macroeconomic situation systematically affecting all observations (See Dwyer (2007)).
From this example, we can see that the 4% default rate has a higher likelihood than the model proposing a 10% default rate, hence Model A would be favored over Model B in terms of a higher chance of observing the actual population default rate of 4% in such model\textsuperscript{14}. At this point, we need to highlight that likelihood metrics are intended to make comparison between models and not to determine if that model is correctly calibrated. For this reason it is necessary to use both: 1) the default rate test to get a clear idea of the agreement between a particular models predictions and the real default probabilities (absolute terms calibration), and also 2) likelihood measures to differentiate which of a set of models makes a better job resembling the real default rates (relative calibration).

As a final note on likelihood measures (applying to calibration in general), we must remember that, when comparing models, we can have a case where we obtain high calibration measures even when the model is poor in terms of power (i.e. a weak model could do best in calibration measures), this of course would be an error. Calibration is complementary to power, there is no guarantee that a model selected only by calibration metrics will the most powerful. This is because a model’s predictions cannot be improved beyond their original accuracy just through calibration unless power is increased first. Again, we have to be careful to start with power as our first model selection criteria.

On other hand, a powerful model will always produce probabilities that are more accurate than a weaker one (even when both are perfectly calibrated). We can see this through the relationship between a model’s power and its predicted probabilities of default. A powerful model, for example, will assign higher PDs for its riskiest obligors (relative to a weaker model) and lower

\textsuperscript{14} In practice, it is recommended to work with the log likelihood to ensure that the biggest likelihood value will be related to the model generating the greatest values. In this case, given that likelihoods are numbers between [0, 1], we will choose the model with the least negative log likelihood as the best calibrated option.
PDs for its safest credits. This will result in flatter calibration curves for weaker models vis-à-vis stronger models.

In the industry, it is sometimes required to calibrate a model’s results to other types of variables. Among the most common alternative calibrations we have: mappings to internal ratings, to historical default rates, and to measures of PD central tendency. In the following lines, we will present briefly, as explained in Bluhm, Overbeck and Wagner (2010), how calibration of PDs to external rating can be done. The first step is to obtain a table with the different credit ratings and their corresponding historic default frequencies. Once we have this information, we proceed as follows:

1. Calculate historical default frequency $h_i(R)$ for each rating class $R$ and each year $i = 1,\ldots, T$. Then calculate the mean value $m(R)$, and standard deviation $\sigma(R)$ of these frequencies. The average $m(R)$ will be our initial guess of the default probability assigned to class $R$, while $\sigma(R)$ will give us an idea of the error we can do when taking the average as an estimate of $R$-class PD.

2. We plot our $m(R)$ values in a coordinate space with rating categories in the horizontal axis, and estimate a (non-parametric) regression curve mapping average default frequencies to ratings. It is noteworthy to mention that there is strong evidence in empirical research that default frequencies increase exponentially with falling creditworthiness.

3. Finally, we use the regression function found in step 2 to estimate the corresponding PDs for each rating category $R$. We can expect that our transformation through the fitted curve will smooth out sampling errors from the historical data.
A more detail exposition of these calibration methods for various types of ratings can be found in Bohn and Stein (2009).

**Estimating Confidence Limits for Power Statistics and Probability estimates**

When developing statistical models, our results are directly affected by the type and quantity of information employed. Therefore, it is crucial to understand to what extent our model’s results have been affected by the particular data set we are using. In the literature this sensitivity is known as *sampling effect*. Given the limited nature of credit information, it is almost impossible for a modeler to know how his model will work when applied to a different sample\(^{15}\), the most he can do is to estimate the change due to sample variability. In the coming lines, we will present some methods to measure the effects of sample variation on power statistics, calibration metrics, and probabilities.

In statistics, we are often faced with the task of predicting the dynamics of events that do not occur very often. One of the main objectives of credit institutions is precisely to avoid outcomes leading to default. As a result, most bank databases contain a reduced number of defaults when compared to the total number of loans. Therefore, while the global amount of information might be important for modeling, it is really the number of defaults what ultimately determines the predictability, stability and quality of our model.

To exemplify this, let us consider a database containing 1,000 loans with a default rate of 5%. A bank would be interested in developing a model discriminating between failing and non-failing

\(^{15}\) As we will see in the Overfitting section, we can split our sample in training, diagnostic and testing sets and/or apply a walk-forward approach to see how our model predicts in a separate (sub-) sample.
loans. Now let us assume that a new observation is added, and that it is a default. In this case, the new default can teach us how it differs uniquely from the other 1,000 non-defaulters. Now what would happen if the new observation is a non-default? In this case, the new observation only yields information on how it differs on only 50 defaulters. As it turns out, a default observation would provide substantially more information than a non-default. We can see that in rare-event problems, as is the case in PD modeling, the number of the scarce observations (default) is what makes a difference.

Another more formal example was presented by Bohn and Stein (2009), where they run a model using randomly selected samples of 50, 100, 200, and 400 defaults and evaluated the variability that each change in sample had on the CAP plot. Each sample size was run one hundred times. The authors present their results in Figure 2.3., where we can see that the smaller the size of defaults in a sample, the larger the variability observed in the CAP plot. The main take away from these exercises is that power statistics tend to vary importantly with the composition of the sample used.

**Figure 2.3. Variability on CAP plots due to changes in the number of defaults in sample**
The literature discusses a number of approaches to determine the variability on power statistics derived from sampling effects (See Delong, Delong and Clarke-Pearson (1988), and Engelman, Hyden, and Tasche (2003). Specifically, Bohn and Stein (2009) provides a closed-form to calculate limits on the sample size needed to estimate and test AUCs and differences between two AUCs, through the following formula:

\[
    n_{\text{max}} = \frac{Z_{\alpha/2}^2}{\epsilon^2} \left[ AUC_1 (1 - AUC_1) + AUC_2 (1 - AUC_2) - 2\rho \sqrt{AUC_1 (1 - AUC_1) \times AUC_2 (1 - AUC_2)} \right]
\]

; where: \( n_{\text{max}} = \text{maximum number of defaults needed to estimate } (AUC_1 - AUC_2) \); \( Z_{\alpha/2} = \text{standard normal variable at the } (1-\alpha/2) \text{ level of confidence; } \rho = \text{correlation between AUCs} \%^{16} \)

\(^{16}\text{According to Bohn & Stein (2009), in PD models, the values of } \rho \text{ are typically in the [0.2, 0.6] range, and can be estimated non-parametrically from a data subset.} \)
In the case we know the number of defaults in the sample \( n \), we can solve for \( \varepsilon \) and estimate the maximum possible error\(^\text{17} \) for \((\text{AUC}_1 - \text{AUC}_2)\) with a \((1-\alpha/2)\) confidence level.

Most importantly to our objectives is the calculation of confidence limits for the predicted probabilities of default. The following procedures will allow us to determine the number of observations needed to conduct accuracy tests on model’s probabilities\(^\text{18} \). In the case when we assume that defaults are independent\(^\text{19} \), we want to make sure that the default rates that we obtained from our model do not differ from the true population’s rates with a confidence level of \((1-\alpha/2)\), formally this can be expressed as:

\[
P(|(d / n) - p < \varepsilon|) \geq (1 - \alpha)
\]

where; \( d = \text{number of defaults in the sample} \); \( n = \text{size of the sample} \); \( p = \text{model’s predicted PD} \); \( \varepsilon = \text{error difference} \); \((1-\alpha) = \text{confidence level}\).

From this equation, we can make sure that the difference between the true and our model’s predicted PD do not exceed a specific amount \( \varepsilon \). On other hand, using this expression and applying the Central Limit Theorem, we can calculate the minimum sample size \( n \) needed to make sure that \( p \) is accurate at a specified \( \alpha \) level, through the following expressions:

\[
P(np_L \leq np \leq np_U) \equiv \Phi \left( \frac{n(p_U - p)}{\sqrt{npq}} \right) - \Phi \left( \frac{n(p_L - p)}{\sqrt{npq}} \right) \Rightarrow 2\Phi \left( \frac{n\varepsilon}{\sqrt{npq}} \right) \geq 1 - \frac{\alpha}{2} \Rightarrow
\]

\[
\left( \frac{n\varepsilon}{\sqrt{npq}} \right) \geq \Phi^{-1}(1 - \alpha/2) \Rightarrow n \geq \frac{pq}{\varepsilon^2} \times (\Phi^{-1}(1 - \alpha/2))^2
\]

\(^\text{17} \) This would give us an idea of the worst scenario (variability between two AUCs).

\(^\text{18} \) For instance, if we require a 99% confidence level and we are dealing with very low default rates, then we would require extremely large number of observations (e.g. for default rate of 50bps, we would need approximately a 33,000 sample).

\(^\text{19} \) And assuming that a binomial distribution will approach a normal distribution as the number of observations increases (Central Limit Theorem). This step is fundamental to derive the size of the sample.
where; \( n = \text{sample size required to make sure that } p \text{ is accurate at } (1-\alpha) \text{ confidence level} \), \( \phi() = \text{standard cumulative normal distribution} \)

Moreover, just as we did above for the AUC differences, we can solve for \( \varepsilon \) to know (for a given sample size \( n \)) how large the difference between our models \( p \) and the empirical default rate \( (d/n) \) needs to be to conclude that this two figures actually are different at a desired level of confidence.

As mentioned above on our discussion of likelihood measures, given that correlation is present in credit information, skewness will appear in the default distribution (i.e. the default rate you observe will be smaller than the expected value and we will observe one large right tail). The presence of this non-diversifiable correlation has important effects on the \( n \) and \( \varepsilon \) we just have calculated. The reason is that the normal approximation of a binomial no longer holds in the presence of correlated default observations\(^{20}\). This causes our estimates to be understated. In the literature, we can find a series of approaches that include correlation in its estimations. Among these we have the analytic solution proposed by Vasicek (1991) were he assumes identical correlations and probabilities across observations, or the simulation of joint factor distributions approach used by Bohn and Stein (2009) where the estimation can be extended to heterogeneous populations with not uniform PDs or correlations. In both works, the main finding are that as correlation increases the estimate of \( \varepsilon \) grows substantially\(^{21}\) given that the right tail (downside) extends and create more extreme values; moreover, given the asymmetric distribution, it

---

\(^{20}\) In the presence of correlation, when we add an observation that is highly correlated to another firm we will not reduce the variability of our portfolio as this new credit will default for sure if the first one defaults.

\(^{21}\) In the simulation approach the \( \varepsilon \) (for a correlation of 0.3) was approximately 3.4 times greater than its equivalent estimate under zero-correlation.
becomes harder to interpret the size of the errors $\varepsilon$ since they occur mainly in the right side extremes.

**Model Robustness: Multicollinearity, Parameter Stability and avoiding Overfitting**

Model robustness is defined as a model’s sensitivity (both of parameters and predictions) to changes in the data. Besides predicting well and accurately, a model’s parameters must be sufficiently stable so that its prediction and relative importance of variables does not change drastically from sample to sample and time to time. Robustness, therefore, complements model’s power and calibration ensuring that they are consistent over time. In this section, we will discuss model robustness from two points of view: multicollinearity and parameter stability.

**Multicollinearity**

In statistics, *multicollinearity* refers to the situation when two or more of the explanatory variables in a multiple regression are highly correlated. This condition distorts the standard error of parameters, leading to issues when conducting tests for statistical significance of factors. The effect of multicollinearity hinders not only the interpretation of the correlated variables, but affects the model’s interpretability as a whole. Since it is difficult to separate the individual influence of each factor, variable selection and model specification is really arduous. But most importantly, the presence of distorted coefficients triggers erratic behaviors in a model’s outputs as small changes in the correlated variables can result in wide variations in PDs predictions.

---

22 In the presence of multicollinearity, the standard errors of the coefficients are “inflated”, this leads to erroneous conclusions about the statistically significance of a variable.
When two independent variables are highly correlated, they basically convey the same information. In some sense, they are competing to explain the same variance in the dependent variable. As a result, the correlated factors will often present coefficients with unexpected or weird signs. For instance, a common phenomenon in regressions with multicollinearity is to see factors with large opposing coefficients (i.e. one of them have a large positive coefficient, while the other shows a large negative one). This occurs because, on one hand, each correlated variable is trying to offset the effect of its counterpart (except for the variance particular to each factor), but as a whole the statistical regression is trying to obtain a combined moderate effect over the dependent (moderate net influence). Given that this two factors are in a continuous tension to explain the dependent variable, we will experience ample variability in our model’s predictions. For example, let us assume that one of the credits has a slightly bigger influence in one of the correlated factors (although the difference might be small), the final result will be that the large (positive or negative) multiplicative effect of that factor will tip the fragile balance into its direction inflating drastically the PD forecast. What we will see is a highly volatile behavior of our model’s output.

So how can we identified if multicollinearity is serious? In the following lines we present briefly two tests to detect high correlation between factors (i.e. VIF and Correlation Matrices); however, there is extensive literature on the topic where additional approaches can be explored (see for example Shmueli 2010, Gujarati 2003, and Greene 1993). The first of these test is known as the Variance Inflation Factor (VIF), which provides a measure of how much the variance of an estimated coefficient is “inflated” because of collinearity. As we will see, the VIF is also defined as the reciprocal of tolerance (e.g. $1-R_i^2$). To present this concept more formally, let us recall the
definition of variance for an estimated coefficient, and defined the VIF as the ratio between the variances of $b_i$ with and without collinearity:

$$\frac{\text{Var}(b_i)}{\text{Var}(b_i)_{\text{min}}} = \left( \frac{\sigma^2}{\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)^2} \times \frac{1}{1 - R_i^2} \right) = \frac{1}{1 - R_i^2} = \text{VIF}$$

; where; $R_i^2 = \text{the } R^2 \text{ value obtained by regressing factor } i \text{ on the remaining independent variables.}$

Ceteris paribus, modelers would want lower levels of VIF, because higher VIFs imply larger standard errors for the parameter estimate. In fact, the convenience of VIF, compared to other measures as *Tolerance*, is that it expressly indicates the magnitude of the inflation in the standard errors of a particular beta due to collinearity. That is: a VIF of 5 indicates that the standard errors are larger five times than would otherwise if there were no correlations between such factor and the rest of explanatory variables in the regression.

It is important to mention that a high VIF does not imply the immediate presence of pairwise correlation (e.g. a variable can also show a high VIF when it is correlated with a *combination* of the other explanatory variables). Therefore a common practice in the industry is to inspect in more detail those independents with VIFs superior to 2\(^{23}\). The VIF analysis should be complemented with other tests, thus, the second suggested method is to look for high pair-wise correlations among regressors in excess of 7.0, conducting an examination of partial correlation coefficients as suggested by Farrar and Glauber (1967). In this research, we will employ Pearson

\(^{23}\) A VIF above 2.0 implies that more than 50% of the variation of that factor is explain by the variation of the other variables in the model.
method to calculate our correlation matrices. Additional tests also include analyzing the condition number \((K = \text{Maximum eigenvalue}/\text{Minimum eigenvalue})\) and condition index \((CI = \sqrt{K})\)^{24}.

**Parameter Stability**

The other aspect of model robustness that we will discuss in this section is parameter stability.

As mentioned above, besides predicting well, a model’s parameters must be sufficiently stable so that its prediction and relative importance of variables does not vary substantially given changes in data used or period of time analyzed. A robust model thus must perform in a consistent way across different data samples.

We must recall that two of the core objectives of statistics are to make inferences about the values of unknown population parameters, and to run hypothesis test about the validity of such values. Typically, the most common hypothesis test is to evaluate if including a factor adds really additional information about the variance of the dependent variable (i.e. assess if a specific coefficient is statistically different to the null hypothesis). In other instances, econometricians might be interested in knowing if a structural change has occurred in a period of time. In such cases, F test are run to evaluate if the ratio of the sum of squared errors of two regressions (one without structural change and the other allowing differences in the coefficients before and after the date of interest) differs statistically from one another. Finally, the model designer might be interested to know if his model’s coefficients will remain consistent if additional or different information is used or if the regression is run over different periods of time. To achieve this, the

---

^{24} According to the rules of thumb presented in Gujarati (2003) to detect multicollinearity: a K between [100, 1000] or CI between [10, 30] would indicate moderate to strong multicollinearity (above 1000 would represent severe MC), while a
simplest method to test parameter stability is to observe the behavior of the parameters over different time windows. If the modeler sees that the size and direction of the factor’s coefficients varies importantly along different periods, the he can deduce that such variable(s) are unstable in the model. In such case, he can eliminate that variable completely from the model or modify the regression specification. On the other hand, if he sees that the parameters remain somewhat constant under (in acceptable statistical variation terms), then he can conclude that these variables and their coefficients are robust and that the model is stable across time.

**Out-of-sample testing: techniques to prevent Overfitting**

In this final part, we will describe a model validation technique that has become a common practice in the credit modeling field: *out-of-sample testing*. Parameters can be highly sensitive to the sample used, hence, to avoid unwanted sample dependency (overfitting), models should be developed and validated using some type of out-of-sample, *out-of-time* testing approach. This validation techniques are highly convenient because they help reduce overfitting of our model on the development data set, and, as we will, see allow us to make use of our information in the most model-convenient and data-efficient way.

As we have mentioned, when developing a model it is convenient to split and reserve a portion of the sample for testing purposes. The initial historical data on which the model is developed is known as the *training* or *in-sample set*, while the data set that has been reserved is known as *test set* or *out-of-sample data*. This procedure is an important element of the evaluation process because it provides a means to assess our model on data that has not been a component of the model definition set. As a result, the model will not have been influenced in any way by the out-
of-sample set and the modeler will have an opportunity to determine how well his definition works on new, different data.

**Figure 2.4. Out-of sample validation techniques**

![Diagram of out-of-sample validation techniques](Source: Sobehart, Keenan and Stein (2000))

Sobehart, Keenan and Stein (2000) present Figure 2.4 where they break up the model testing process across two main dimensions: time and borrower’s population. The upper-left section presents the case in which testing data is chosen completely at random from the full sample. As mentioned by the authors, this approach assumes that data follows a stationary process. Moreover, given that the testing set is selected randomly, this selection method validates the model along the obligors dimension but does not check for variance across time.

On other hand, the upper-right section presents the most common approach for the selection of the testing set. In this case, training data is selected from any time period previous to a certain date and testing data is picked only from periods after that date (e.g. training set using data from 2000-2005, and testing set using 2006-2008 data). This *out-of-time* procedures allows the detection of time variability, but assumes that the population does not vary across different obligors.
In the bottom-left section the data is split in two sets that have no obligors in common. When the population for the training set is completely different than the one for the testing set, the validation would be out-of-the-universe (e.g. a model is trained on telecommunication sector but tested on energy firms). This technique will not detect time dependence in data but will detect changes across population.

Finally, the bottom-right section presents a segmentation procedure where data is split both through time and across population. For example, in this approach we would train our model using all credits on a particular segment A from 1990 to 1995, and tested on a sample of low-scored loans for segment B from 1996 to 2000. This should be the preferred sampling method.

As we have seen, using out-of-sample/time techniques, without doubt, is the best way to evaluate a model’s performance. However, given that default information is so scarce, modelers have to be very careful on how they split their data between development and training sets. On one hand, if too many failing credits are excluded from the training set, estimation power will be seriously affected (as we saw on our discussion of variability of CAP plots). Conversely, if too many defaulters are omitted from the testing set we will not be able to truly evaluate the consistency of our model in samples different than the one we are using.

Given these issues, a walk-forward testing\textsuperscript{25} approach can be applied to make the most efficient use of our available data by combining out-of-time and out-of-sample tests (bottom-right section in Figure 2.4). Unlike in-sample measures of fitness (e.g. $R^2$, or AICs)\textsuperscript{26}, walk-forward testing

\textsuperscript{25} This term is borrowed from the trading model literature, see for example Pardo, Robert (2008); “The Evaluation and Optimization of Trading Strategies”; 2\textsuperscript{nd} Edition, Wiley.

\textsuperscript{26} Which are designed to measure the fit of a model's parameters or errors relative to the development data set.
allows us to evaluate the predictive power of our model when applied to data different to its training set. The steps as described in originally in Sobehart, Keenan and Stein (2000) are:

1. Choose a year (e.g. 2009), and regress the model with all available data from and before that particular year (i.e. 2009, 2008, 2007...).

2. Once the model has been defined, generate predictions for all the loans available during the following year (i.e. 2010). Save these predictions as part of a result set. Notice that the predictions we just generated are out-of-time for firms present in previous years and out-of-sample for all firms available from 2009 onwards.

3. Move up the window one more year (e.g. 2010) and use all available data from and before that year to re-train the model (i.e. 2010, 2009, 2008...), and use the data for the next year (i.e. 2011) for testing.

4. The process is repeated from steps 2 to 3 for every year, adding the new predictions to the result set each time.

The result set (containing out-of-sample, out-of-time data) can then be employed to test the performance of the model in more detail. It is important to note that the validation set is a sub-set of the population and can yield spurious results as result of data anomalies. In this case, resampling techniques can be implemented. An example of a useful resampling technique is: 1) select a sub-sample from the result set, 2) calculate a performance measure for the sub-sample (e.g. sensitivity), record that value, 3) select another sub-sample, and repeat the process; this continues until a distribution of the performance measure is obtained. This process provide an estimate of the variability of the actual model performance. Moreover, in cases where the distribution converges to a known form, this information can be utilized to assess if differences in model performance are statistically significant.
As we can see, walk-forward testing is convenient because: 1) it provides modelers an iterative opportunity to assess the performance of the model across time, and 2) gives an effective means to reduce the risk of in-sample overfitting by leveraging the availability of data in the most efficient way.

2.4. Transforming Statistical insight into Economic terms

As mentioned above, one of the premises of this work is to demonstrate that the benefits of statistical models are not confined to improved prediction accuracy but that they can be translated into concrete economic and operative returns. Moreover, we are convinced that the gains from quantitative-supported lending processes will result not only on better cost-benefit structures and market targeting for financial intermediaries, but will be passed through to low-income customers in the form of fairer interest costs and extended access to credit.

Optimal Cutoff Estimation

As explained by Bohn and Stein (2009), given that business decisions are more easily understood in terms of money and savings rather than in ROC terms or statistical precision, modelers must be able to convert these measures into clear estimates of profitability. In the following lines, we will explore how to determine the optimal cutoff that minimizes costs through the use of ROC curves.

A loan threshold or cutoff point is defined as a PD level or risk score below which a bank will not concede loans. In the industry, cutoffs are often used to provide an initial indicator against which to evaluate credit applicants, but these thresholds can be the base for more precise lending policies with detailed information about a company’s cost structure and prior probabilities.
As discussed in the confusion matrices section, a different threshold will change the number of observation in each types of errors. Thus, different cutoffs will generate different observations in each category. Our definition of an optimal cutoffs relies on this fact and tries to find the PD level that optimizes the total economic payoff associated with that strategy. To do this, a strategy’s total payoff is defined as a function of the benefits related with a correct prediction (True Positive and True Negative), the costs resulting from Type Errors I and II (False Negative and False Positive), and the overall performance of the model\textsuperscript{27}. Under these premises, the higher the error rate of a model, the more costly will be to use it.

Following the definition presented in Bohn and Stein (2009), the total payoff of using a specific PD model \(m\) and a cutoff \(k\) is given by the expected benefits of correct predictions minus the expected costs of type errors I and II associated with a particular threshold \(k\). Formally the expression is:

\[
Total\ Payoff_{m,k} = \left[ p(D) \cdot b(TP) \cdot TP_{m,k} + p(ND) \cdot b(TN) \cdot TN_{m,k} \right] \\
- \left[ p(D) \cdot c(FN) \cdot FN_{m,k} + p(ND) \cdot c(FP) \cdot FP_{m,k} \right] 
\]

where: \(p(\cdot)\) = unconditional probability of an event; \(b(\cdot)\) = benefit function; \(c(\cdot)\) = cost function; \(D\) = default event, \(ND\) = Non-default event;

\textsuperscript{27} It is important to note that, in this analysis, the costs and benefits are assumed to be independent of the model used.
From our past definition and based on Figure 2.5, we can re-express the total payoff as a function of corrected predictions and specific types of errors (I and II) in the following way:

\[
\text{Total Payoff}_{m,k} = [p(D) \cdot b(TP) \cdot ROC(k) \perp p(ND) \cdot b(TN) \cdot (1 - k)]
- [p(D) \cdot c(FN) \cdot (1 - ROC(k)) + p(ND) \cdot c(FP) \cdot k]
\]

Now, to identify the optimal cutoff point we need to differentiate this total payoff function with respect to \(k\) and then setting this expression equal to zero. Formally, the expression would be:

\[
S = \frac{dROC(k)}{dk} = \frac{p(ND) \cdot [c(FP) + b(TN)]}{p(D) \cdot [c(FN) + b(TP)]}
\]

The result is the slope of a line where marginal costs and marginal benefits are equal (marginal total payoff equals zero). The tangency point between this line with slope \(S\), known as *Iso-
performance line, and our model’s ROC curve will define the optimal threshold \((k^*)\) given the company’s specific cost, benefit functions. In the literature\(^{28}\) it is demonstrated that at this tangency point both Type Errors I and II are minimized.

As discussed above in the model’s power section, we know that when the ROC of a model strictly dominates another, that model will provide better discriminatory power. Hence, under our present argument, that dominant model will be preferred over any threshold. On the other hand when we have two models whose ROCs intersect, then no model will unequivocally dominate the other, thus the preferred model will depend on the type of borrowers we are targeting \textit{ergo} the cost function utilized.

**Power measures and Positive Net Present Value lending**

An alternative approach is to use the conditional probabilities from our models to determine estimates of profitability based on a positive net present value (NPV) criteria. In many instances, credit managers decide to make a loan whenever the net present value of expected cash flows from a credit is positive. In formal terms we have:

\[
NPV = \pi * V_D + (1 - \pi) * V_{ND}
\]

; where: \(\pi = \text{unconditional (population) probability of default}; \ V_D = \text{payoff value under default}; \ V_{ND} = \text{payoff value in non-default}\)

Reformulating this expression in terms of correct predictions, types of errors, PDs and cost-benefit functions we obtain the following expression:

NPV = \[ p(D) \times b(TP) - p(D) \times c(FN) \] + \[ p(ND) \times b(TN) - p(ND) \times c(FP) \]

where: \( \pi = p(D) \), and \( 1-\pi = p(ND) \); \( V_D = b(TP) - c(FN) \); and \( V_{ND} = b(TN) - c(FP) \)

By substituting our model’s conditional probabilities and assuming a constant cost-benefit function, the past NPV expression can be reformulated as:

NPV = \[ p(D) \times b(TP) \times TP_{m,k} + p(ND) \times b(TN) \times TN_{m,k} \]
- \[ p(D) \times c(FN) \times FN_{m,k} + p(ND) \times c(FP) \times FP_{m,k} \]

We can see that this last equation is just the negative of the Total Payoff expression we obtained above. This implies that setting our threshold at the point where a portfolio’s marginal NPV is zero would be equivalent to establishing our decision threshold in the point where the bank’s marginal costs are minimized (tangency point between ROC and S line above). Under this finding, the bank should continue to give credits to the point where marginal returns become zero, this would provide at the same time the strategy minimizing costs for that specific model \( m \) and threshold \( k \).

As mentioned above, the definition of optimal cutoffs (from a cost or benefit perspective) based on the results of credit scoring models can help credit managers to estimate the benefits of credit concession strategies in a more precise way. For example, by introducing additional benefits related to a certain loan strategy (i.e. lending to bank clients who also generate relationship benefits) in our NPV/Total Payoff expression, we would obtain a steeper Iso-performance curve \( S \), shifting our optimal cutoff \( k \) to the left. This smaller cutoff for relationship clients should not be understood as preferential treatment per se, but as the consequence of the higher relative benefits when preferential clients do not default (i.e. additional revenue from adjacent
businesses). This is a clear example of the use of statistical models as a guide for informed decision policies in the credit concession field.

**Optimal Pricing Estimation, Zero-cost lending and Capacity restrictions.**

In the past section, we focused on the costs and benefits for *all* obligors relative to a specific cutoff (determine threshold given a cost function). In this part, we will determine how the marginal revenue can be adjusted to compensate for the risks and costs associated with a predetermined lending objective (i.e. optimal pricing).

To understand this, let us first analyze the case where a financial institution has limited capacity and can only process a specific number of applicants. This example will help us understand the dynamics when the cutoff is given. In this case the financial institution will adjust its lending terms to the particular cutoff to which it is limited\(^{29}\). That is, given a *pre-specified* threshold, the lender will determine the minimum level of marginal benefits at each risk level that will ensure zero average costs for lending at that \(k^*\). Formally, this means setting our total payoff function to zero, and solving for the amount of revenues \(b_{k=k^*}(TN)\) that counterweights the costs associated with our model\(^{30}\). This yields the following expression:

\[
b_{k^*}(TN) = \frac{1}{p(ND) * (1-FP)} \left[ p(D) * c(FN) * FN_{m,k} \perp p(ND) * c(FP) * FP_{m,k} \right]
\]

---

\(^{29}\) Unlike the past case in which we determined the optimal cutoff given a pre-specified costs-benefit structure.

\(^{30}\) It is noteworthy to mention that in the coming expression we are considering null benefits from failing borrowers who were denied a credit (i.e. \(b(TP=0)\)).
It is important to mention that \( b_{k=k^*}(TN) \) is the marginal revenue required to transform the pre-specified cutoff \( k \) into the optimal cutoff \( k^* \) (i.e. the level of benefit that needs to be received from every obligor approved\(^{31}\)).

This same approach can be extended to estimate loan prices. In the case of pricing, our objective is to determine an appropriate level of revenue for a loan with score \( k \), under the assumption that no credit will be denied. Given that all loans will be conceded, the benefit calculation will be defined in terms of only TNs and FNs, provided we determine the corresponding adequate \( b_k(TN) \) at each score \( s(k) \) (i.e. we are charging a different price for each loan class). Under these assumptions our original NPV maximization condition would be reformulated as:

\[
\frac{dNPV}{dk} = p(ND) \times b(TN) \frac{\partial TN}{dk} - p(D) \times c(FN) \times \frac{\partial FN}{dk} = 0
\]

Solving for \( b_k(TN) \), where the subscript indicates that this is the level of marginal benefits (price) that should be charged according to the specific costs/risks related to each score \( s(k) \), we obtain the following break-even price expression:

\[
b_k(TN) = \frac{p(D) \times c(FN) \times \frac{\partial FN}{dk}}{p(ND) \times b(TN) \times \frac{\partial TN}{dk}}
\]

It is noteworthy to mention that this price expression, unlike the cutoff estimation case, calculates a revenue in terms of the specific costs associated with scores equal to a particular point in the ROC curve, and not based on cumulative (and uniformly assigned) costs across all borrowers. In this way, by calculating the corresponding level of profit at each value of \( k \), lenders would be able to derive a pricing curve for all levels of credit quality. That is, creditors can concede loans to any class of borrower provided the corresponding adjustment to the risk profile.

\(^{31}\) This assumes that every applicant receives a loan if his score is above \( k \) (at a uniform price level), and no applicant with score below \( k \) receives a loan.
each obligor. On other hand, it is important to highlight the importance of robust and powerful models in estimating appropriate prices, as weaker models will consistently misestimate (under and overprice) the level of revenue that should be charged to customers with lower and higher credit score classifications. This demonstrates that powerful discriminatory models can represent a strategic advantage for lenders.

In the next chapter we will adapt all of these approaches to a consumer lending provider in the low-income segment. By applying these quantitative tools to a real context, our objective is to demonstrate and quantify the direct benefits of using accurate PD estimation techniques to influence the decision process, and prove to lenders in the base-of-the-pyramid unfamiliar with statistical jargon that PD models can be immediately translated into clear business terms and conventional lending practices. Our aim is to show that these two world are complementary rather than contradictory.

CHAPTER 3. MODEL DEVELOPMENT AND RESULTS
“Statistical thinking will be one day as necessary for efficient citizenship as the ability to read and write”

H.G. Wells (1952)

3.1. Data provider: MIMONI

Our research is based on a 2,000 loan sample provided by Mimoni Group. Mimoni is a pioneer in the use of data analytics applied to low-income consumer-lending segments in Latin American. As mentioned in our Introduction, the lending industry in Mexico requires long and invasive processes, unfair collateral requirements, and high interest rates to obtain a credit authorization, extending the financial exclusion of the base-of-the-pyramid. Through the use of quantitative models, Mimoni provides an online credit application for short-term unsecured loans ranging from $1,000MX to $4,000MX at a 0.25% daily interest rate and a maturity ranging from 1 to 5 fortnights. Mimoni’s credit risk assessment algorithm requires no face-to-face, endorsement, or collateral, speeding-up and easing the loan granting process. Through its online application clients can obtain a credit authorization in just minutes. As long as clients make payments on time, their credit terms improve having access to higher amounts, longer payment periods, and preferential (lower) interest rates. Currently Mimoni is operating in Mexico City and has started to expand its operations at a national scale and some parts of US.

Mimoni Group, formerly known as MiCel, was founded in 2008 by MIT student Pedro Zayas, and Stanford MBA Gabriel Manjarrez with the support of two of the largest global impact investment firms (IGNIA and Omidyar Network) and two Venture Capital firms based on Silicon Valley (Storm Ventures and Bay Partners). Through the use of proprietary credit scoring processes and algorithms, Mimoni discriminates among applicants defining their creditworthiness level based on the credit behavior and payment patterns of a 50,000+ database,
using data analytics based on socio-economic borrower’s profiles, credit bureau information and other proprietary cross-validation sources. Mimoni is Mexico’s first online short-term loan provider, enabling almost instant-approval lending for tens of millions of credit-worthy clients that today are financially under-banked, and have to rely on the unfair lending terms of money lenders and pawn shops.

Mimoni credit application process: how it works

To apply for a loan, the user needs to access and register in Mimoni’s website http://www.mimoni.com/Index.aspx and fill out the credit evaluation form with requires personal information (i.e. age, number of dependents, marital status, etc.), the amount she is applying to, as well as a payment time that ranges between 1 to 5 fortnights. This information is processed into the system and cross-validated with credit bureau information (if available), creating a socio-economic profile and credit risk scoring for the borrower.

Based on this information, the applicant is notified in minutes about the authorization or decline of its credit. If approved the loan amount is deposited on the bank account indicated by the client (or on a debit card) in less than 24 hours. When the loan is approved for the first time, the client is required to sign a contract in person (whether in Mimoni’s offices or in the client’s address). It is important to mention that if the client paid on time on past loans, he is automatically approved for consecutive credits. The loan applicant needs to: 1) to be 18+ years old, 2) have an official ID; and 3) own an e-mail address.
3.2. Database description

Our original data base is comprised of 2,000 short-term loans with a composition of 37% of defaulting credits and 63% non-defaulting loans (i.e. 741 failing and 1259 non-failing respectively). Each credit in the sample was made for a $1,000MX loan amount, and an expected final payment of $1,200MX (i.e. 20% interest rate over a 15 days period).

Each observation contains information on 22 different variables across three main categories: 1) Demographic Information, 2) Credit Bureau history, and 3) Loan specific information. Many of these variables were transformed and redefined to fit our modeling purposes. For instance, we developed 7 additional binary and multiclass variables to capture different behaviors on our dependent variable and a set of dichotomous for most of our categorical factors (further detail and the rationale behind this transformations is presented below). Table 3.1 presents a dictionary with the description of the original variables provided by Mimoni and the additional ones we generated for our regressions.

Table 3.1. Variables Description

<table>
<thead>
<tr>
<th>NAME</th>
<th>DESCRIPTION</th>
<th># OF LEVELS</th>
<th>LEVELS</th>
<th>VARIABLE CATEGORY</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Sex</td>
<td>Categorical variable on the obligor's gender</td>
<td>2 levels</td>
<td>F= female; M= Male</td>
<td>Demographic</td>
</tr>
<tr>
<td>2 Age</td>
<td>Numerical variable on the borrower's age in</td>
<td>num</td>
<td>e.g.: 31.6 years</td>
<td>Demographic</td>
</tr>
<tr>
<td></td>
<td>fractions of a year</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 SpouseAge</td>
<td>Numerical variable on the borrower's spouse's</td>
<td>int</td>
<td>e.g.: 51 years</td>
<td>Demographic</td>
</tr>
<tr>
<td></td>
<td>age in years</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Education</td>
<td>Categorical variable on borrower's level of</td>
<td>6 levels</td>
<td>Basic; Jr. High; High-school;</td>
<td>Demographic</td>
</tr>
<tr>
<td></td>
<td>education</td>
<td></td>
<td>Technical; Bachelor; Postgrad.</td>
<td></td>
</tr>
<tr>
<td>5 MaritalStatus</td>
<td>Categorical variable on borrower's marital</td>
<td>6 levels</td>
<td>Single; Married;</td>
<td>Demographic</td>
</tr>
<tr>
<td></td>
<td>status</td>
<td></td>
<td>Separated; Divorced;</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Widowed; Living together</td>
<td></td>
</tr>
<tr>
<td>6 Dependents</td>
<td>Numerical variable on the number of people</td>
<td>int</td>
<td>e.g.: 2 people</td>
<td>Demographic</td>
</tr>
<tr>
<td></td>
<td>directly depending on the borrower</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 HomeOwnership</td>
<td>Categorical variable on borrower's house</td>
<td>4 levels</td>
<td>Own house; Rented;</td>
<td>Demographic</td>
</tr>
<tr>
<td></td>
<td>ownership status</td>
<td></td>
<td>Mortgaged; Family own</td>
<td></td>
</tr>
<tr>
<td>8 MonthsInCurrent</td>
<td>Numerical variable on the number months</td>
<td>num</td>
<td>e.g.: 4.5 months</td>
<td>Demographic</td>
</tr>
<tr>
<td>Residence</td>
<td>living in current residence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 Employment</td>
<td>Categorical variable on borrower's employment</td>
<td>26 levels</td>
<td>e.g.: Agriculture;</td>
<td>Demographic</td>
</tr>
<tr>
<td></td>
<td>sector</td>
<td></td>
<td>Construction;</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Consulting; Government;</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Independent; Manufacturing;</td>
<td></td>
</tr>
<tr>
<td>10 MonthsInCurrent</td>
<td>Numerical variable on the number months</td>
<td>num</td>
<td>e.g.: 4.62 months</td>
<td>Demographic</td>
</tr>
<tr>
<td>Employment</td>
<td>living in current job</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No.</td>
<td>Variable Name</td>
<td>Description</td>
<td>Type</td>
<td>Example</td>
</tr>
<tr>
<td>-----</td>
<td>-----------------------------------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>----------</td>
<td>--------------------</td>
</tr>
<tr>
<td>11</td>
<td>TotalDeclaredIncome</td>
<td>Numerical variable on borrower's declared Income per month (in MX Pesos)</td>
<td>int</td>
<td>e.g.: $7,200</td>
</tr>
<tr>
<td>12</td>
<td>TotalMonthlyPayments</td>
<td>Numerical variable on borrower's total debt commitments (Includes past debts registered in the Credit Bureau (see row 14) + other additional debts reported by borrower); 0= borrower reported zero debts with other institutions; The difference between row 12 and 14 would give additional debts besides those in Credit Bureau (see below row 30)</td>
<td>int</td>
<td>e.g.: $13,625</td>
</tr>
<tr>
<td>13</td>
<td>CreditHistoryLengthMonths</td>
<td>Numerical variable on borrower's total amount of time with records in the Credit Bureau in months; 0= implies borrower does not have a record in Credit Bureau</td>
<td>num</td>
<td>e.g. 3.5 months</td>
</tr>
<tr>
<td>14</td>
<td>DebitPaymentsMonthly</td>
<td>Numerical variable on borrower's Accumulated amount of debts as recorded in Credit Bureau (Closely related to row 12, depending on the accuracy of debt declared by borrower); RA= borrower does not have a record in the Credit Bureau</td>
<td>Int</td>
<td>e.g.: $27,422</td>
</tr>
<tr>
<td>15</td>
<td>RevolvingCreditAuthorized</td>
<td>Numerical variable on the Maximum amount of credit authorized in ALL of borrower's credit cards (i.e. if borrower has 2 credit cards this figure presents the sum of credit in both cards although issued by different institutions); 0= borrower possess no credit cards, or has 0 authorized credit</td>
<td>Int</td>
<td>e.g.: $15,383</td>
</tr>
<tr>
<td>16</td>
<td>RevolvingCreditUsed</td>
<td>Numerical variable on the total number of credit actually used (i.e. sum of credit used in all borrower's credit cards), if row 15 &gt; row 14 implies borrower has overdrawn his authorized credit</td>
<td>Int</td>
<td>e.g.: $5,143</td>
</tr>
<tr>
<td>17</td>
<td>MaximumCreditLimit</td>
<td>Numerical Value with the highest historical Maximum credit authorized in one of the borrower's credit cards (It does not have to match the revolving/current figures)</td>
<td>Int</td>
<td>e.g. $6,000</td>
</tr>
<tr>
<td>18</td>
<td>LoanAmount</td>
<td>Numerical Value on the size of the loan granted (ALL observations have the same amount = $1,000MX)</td>
<td>Int</td>
<td>$1,000</td>
</tr>
<tr>
<td>19</td>
<td>LoanTotalDue</td>
<td>Numerical Value on the size of the loan granted (ALL observations have the same amount = $1,000MX)</td>
<td>Int</td>
<td>$1,200</td>
</tr>
<tr>
<td>20</td>
<td>LoanDueDate</td>
<td>Original loan due date at the beginning of loan concession (not renegotiated due date)</td>
<td>date</td>
<td>e.g.: 8/16/2003</td>
</tr>
<tr>
<td>21</td>
<td>FirstPaymentDelay</td>
<td>Numerical Value on the number of days of delayed payment relative to the original loan due date; 30+ days delay = Past due; 90+ days delay = Full Payment Delay → Borrower paid the full amount on same date.</td>
<td>num</td>
<td>e.g. 95 days</td>
</tr>
<tr>
<td>22</td>
<td>FullPaymentDelay</td>
<td>Numerical Value on the number of days of delayed payment relative to original due date. Borrower needs to pay at least $300MX to have an extension. Just as before: 30+ days delay = Past due; 90+ days delay = Uncollectibles; NULL= Defaulted (never paid); If FirstPaymentDelay = Full PaymentDelay → Borrower paid the full amount on same date.</td>
<td>num</td>
<td>e.g. 15 days</td>
</tr>
<tr>
<td>23</td>
<td>Class1</td>
<td>Multi-class variable derived from Full Payment Delay (row 22)</td>
<td>4 levels</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>BinFull_RA</td>
<td>Binary variable derived from Full Payment Delay column (row 22) based on a Broad &quot;Risk Averse&quot; definition of Default [i.e. F1 (loan never paid) + F2 (Uncollectibles) + F3 (Past Due)]</td>
<td>2 levels</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>BinFull_RP</td>
<td>Binary variable derived from Full Payment Delay column (row 22) based on a shorter, &quot;Risk Prone&quot; definition of Default [i.e. F1 (loan never paid) + F2 (Uncollectibles)]</td>
<td>2 levels</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>ClassPF_RA</td>
<td>Multi-class variable derived by the combined behavior of First Payment delay (row 21) and Full Payment Delay (row 22), for example (10, NULL) implies that a borrower was delay on original due date by 10 days, and then he never paid after the extension; &quot;Risk Averse&quot; definition (broader).</td>
<td>4 levels</td>
<td></td>
</tr>
</tbody>
</table>
As mentioned above, we developed 7 additional binary and multiclass variables to capture different aspects in the loan default behavior. To achieve this, we defined first four subcategories of borrowers from those with the less to those with the most desirable payment pattern. Such classification serves two purposes, first, it allows us to identified granular default behaviors using multi-class regressions, and second is the base for the construction of further dependent variables definitions under risk averse or risk prone lending strategies. The pre-classification is as follows:

1. **Strict Defaulter.** Defined as individuals that received a loan and never paid it back. This category is classified in most researches as the prototypical default event. It is noteworthy to mention that in our sample we identified several cases in which the borrower paid the minimum amount of $300MX to renegotiate the due date. Under all
our different dependent definitions, we considered that loan to have defaulted as the lender never received full completion of the payment at any date.

2. **Uncollectible loans.** Given the business definitions provided by Mimoni, any credit with a payment delay of more than 90 days is considered as lost fund. Most credit lenders tend to include (a priori) this category in their default event definition given that they presume this loan will never be collected. Hence, most of our alternative dependent variables include this category under the failing credit category.

3. **Past due loans.** Following the industry practice, Mimoni defines any loan with more than 30 but less than 30 days of delay as a delinquent (although) collectible loan. As we will mention below, a risk averse lender will be inclined to include this category under the default event, although strictly speaking this loan is still collectible. Most risk prone institutions would include this type of loans in its non-failing category.

4. **Regular loans.** Any loan with less than 30 days delay (including those on or before due date) as classified as regular loans. It would be interesting to explore the behavior of those borrowers with prompt payment behavior (i.e. early payers); however, the number of observations in this category was too small for more granular analysis.

As recommended by Mimoni’s top management, we decided to explore several definitions of what constitutes a good and bad borrower, which will depend on the risk/return criteria used by different lenders. In our case, we defined lenders under the two following lending profiles:

a) **Risk Averse lending approach.** Under this strategy the lender is more interested in avoiding the default event; therefore, a broader definition of failing loan should be used. For instance, on this type of lending practice, a lender could include “Strict Defaulters”,

84
“Uncollectible loans”, and (depending on the degree of aversion) “Past Dues” to include any loan with delinquent behavior.

b) **Risk prone approach.** Under this approach, a financial institution can be more flexible about what constitutes a delinquent loan and be more interested in the possible revenues derived from a more relaxed lending strategy (i.e. “word of mouth” positive externalities, rate of collection of past dues). As we explored in Chapter 2, a lending institution might be willing to grant more loans provided a price that compensates for the additional risks/costs associated with a borrowers’ type. Thus, although a loan might be overdue by 30 days, a lender could more than compensate its costs by charging a price rewarding for the extra risk. As we saw in Chapter 2, one way to determine this appropriate price is by deriving our NPV function relative to the k-customer specific cutoff and solving for \( b(TN) \). All of these decision depend on the lender’s specific cost-benefit function and accuracy of its discriminatory model (e.g. Logit PD model).

Using this tow lending criteria, we generated the seven different versions of our dependent variable including a multi-class version, and two binaries (risk averse, and risk prone) based on the final re-negotiated payment date (i.e. “FullPaymentDelay”); and two multiclass variables (one reflecting risk aversion and other under more risk inclination), and its corresponding dichotomous versions to capture the combined behavior of delay under the original and extended payment date (i.e. “FirstPaymentDelay” and “FullPaymentDelay” variables in dictionary).

---

3.3. Data set cleaning, and variable transformations

In the following section, we present the different variable transformation procedures that we undertook to ensure maximum usability of our original data sample. Given that six out of our 23 independent variables where categorical, we had to generate a series of binaries to capture the explanatory power of the different levels within each variable. The binary variables we constructed based on the number of observations in each level and under the following criteria:

1. **Education.** As detailed above in our variable dictionary, the Education category is comprised by 6 levels including: Primary Ed.; Jr. High; High-school; Technical; Bachelor; Postgrad. To simplify the number of variables we grouped these levels in three main binaries: Basic Education (= Primary + Jr. High); Medium (= High School + Technical Education); and High Education (Bachelor degree + Post. Grad).

2. **Marital Status.** As in the case above, our objective was to cluster the number of levels in as less binaries as possible. The binaries defined are: Singles (formed a single group given the large number of observations in this category); Couples (=Married + Living Together); Head of Household (= Separated + Divorced + Widowed).

3. **Homeownership.** These classifications were built based on the number of observations in each level and on the premise that a borrower owning his home has more financial flexibility than those with some sort of fixed housing payment, and more financial power than one living with familiars. Therefore the four levels were group in: Own; Obligation (= Renting + Mortgaged); Family House.
4. **Employment.** This variable contained 26 levels which we grouped in three main categories including: Independents/Other activity (with 901 observations); High Value Employment (including those categories with typically higher incomes given the type of activity, e.g. Medical Services, Consulting, Telecom, etc.); and Low Value Employment (comprising those activities with typically low incomes in Mexico, e.g. Agriculture, Construction, etc.).

5. **Loan Due Date.** Given that a time-related variable can potentially yield information about the seasonality and cyclicity in payment behavior (e.g. default events can trigger on period when people has more expenses as in December), we generated a 2 levels factor from our original Loan Due Date variable (‘DATE’ type). The levels are: 1 (= winter and summer months); 0 (= otherwise).

6. **Payments Declared and Is_In_Credit_Bureau variable.** Given that the information in Total Monthly Payments (row 12) and Debt Payment Monthly (including the Debts registered in the Credit bureau; row 14) is closely related, we decided to generate two additional variables: one capturing the additional reported expenses besides those reported Credit Bureau, called “Payments Declared”, and a second binary variable measuring whether a borrower has a record in the Credit Bureau. As we will see below, the idea is to capture as most causalities as possible, while choosing those variables with highest explanatory power (and less linearly correlated) through a k-step regression process (See regression section below).

7. **Income-Debt ratio.** Finally we calculated a measure of leverage by borrower relative to their level of income. In many PD default models developed for large and small companies, leverage is a powerful explanatory variable. Assuming this causality is
maintained for smaller credit entities (i.e. consumer borrowers), we included such variable in our data set.

**Multivariate Imputation by Chained Equations**

As mentioned above, we presume that Credit Bureau information can have a significant explanatory power in probability of default prediction. Therefore, to take the most out of our existing sample we made use of *Multivariate Imputation by Chained Equations* (MICE) to compensate for missing variable (NAs) in our Total Monthly (row 12), and Debt Payments Monthly (row 14) variables.

Imputation can take one of two general approaches; Joint Modeling and Fully Conditional Specification. Joint Modeling is useful when a suitable multivariate distribution is found for the data. Thus, Joint Modeling specifies a multivariate distribution for missing information and calculates imputation from their conditional distributions. Fully Conditional Specification, on other hand, is preferred when no reasonable multivariate distribution is found. The method then works specifying a multivariate imputation model for each variable based on a set of conditional densities, and running a small number of iteration (starting from an initial iteration) until valid data is reached. Our main driver by applying MICE to our data is that we want to be able to predict for a representative sample of borrowers, rather than just the ones with all data reported, avoiding thus possible bias on our results.

**Log Transformations**

As can be seen in Figure 3.1.A, seven of our variables in the original data set present a skewed distribution. These variables are: Revolving Credit Used; Revolving Credit Authorized; Maximum Credit Limit; Debt Payments Monthly; Total Declared Payments; Total Income
Declared, and Income Debt Ratio. Given the importance of these variables over the dependent variable, we proceed to run a logarithmic transformation to minimize the effects of outliers, obtaining observations approximately symmetrically distributed. The results of such transformation are shown in Figure 3.1.B.

**Figure 3.1.A. Skewed variable distributions before logarithmic transformations**

![Histograms showing skewed distributions](image1)

**Figure 3.1.B. Symmetrical distributions after logarithmic transformation**
Correlation Matrices

As mentioned in Chapter 2, multicollinearity can create significant distortions in our regression affecting not only the correlated variables but the interpretability of the model as a whole. As we explained there, when two explanatory variables are highly correlated, they basically compete to explain the same variance in the dependent.

Multicollinearity occurs when a given predictor can be approximated by a linear combination of the other variables in the model. This has the effect of increasing the variance our coefficients. As a result, the correlated factors will present unexpected or counterintuitive signs. Hence, before running our regression, we conducted two tests to identify the existence of highly correlated variables, these test are: Variance Inflation Factor (VIF), and Correlation Matrices.
As explained above, VIF provides a measure of how much the variance of an estimated regression coefficient is inflated because of collinearity. Therefore, we look for lower levels of VIF as higher values would imply larger standard errors. In practice, we tend to eliminate variables with VIFs around 5.

Table 3.3 presents the VIF values generated by regressing our independent against all the explanatory variables including the logarithmic transformations for the variables included in Figure 3.1. As we can see, only Income Declared (4.5876) and Payments Declared (6.6554) present high VIFs, which would indicate significant correlation with other variables. It is noteworthy to mention that, by eliminating Payments Declared from the regression, Income Declared returns to normal VIF levels of 1.419451 (See Appendix A.1.1 for recalculated figures).

Additionally to VIF, we also calculated Correlation Matrices using Pearson’s method. It is important to highlight that, given that we run Logistic regressions using levels and Log transformations of our independent variables, we generated correlation matrices for both cases. For our analysis, we classified correlations between [0.55, 0.77] as medium-high, and any value above 7.8 as a high correlation event. Our results are summarized in Tables 3.4 and 3.5.

As we can see from Table 3.4, we have only two cases of high correlation that could represent a problem in our regressions. In the first case, the Income Declared variable shows an important degree of correlation with Payments Declared (i.e. 0.637438), a result that is consistent with our VIF findings. On the other hand, Revolving Credit Authorized and Revolving Credit Used present a serious problem of pairwise correlation with a 0.785803 figure, indicating that we will have to evaluate which of these variables has to be eliminated from our regressions.
Table 3.5, on other hand, present our Correlation Matrix including logarithmic transformations for seven of our independent variables. As we can observe, *Income Declared* and *Payments Declared* show a similar level of correlation level as the one observed using levels (0.6175262). In the same line, *Revolving Credit Authorized* and *Revolving Credit Used* also display a similar high correlation around 0.74.

However, in this case, the Log transformations allowed to identify an additional correlation case between *Payments Declared* and *Income Debt Ratio*. This findings, coupled with our results above, would suggest that *Payments Declared* is a good candidate to be eliminated from our models given its high VIF figures and high correlation levels with two other variables.

### Table 3.3. VIFs on Logistic regression including all independent variables (Log transformations)*

<table>
<thead>
<tr>
<th></th>
<th>GVIF</th>
<th>Df</th>
<th>GVIF*(1/(2*Df))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>1.491432</td>
<td>1</td>
<td>1.221242</td>
</tr>
<tr>
<td>Dependents</td>
<td>1.290760</td>
<td>1</td>
<td>1.136116</td>
</tr>
<tr>
<td>MonthsInCurrentResidence</td>
<td>1.233191</td>
<td>1</td>
<td>1.110491</td>
</tr>
<tr>
<td>MonthsInCurrentEmployment</td>
<td>1.268801</td>
<td>1</td>
<td>1.126411</td>
</tr>
<tr>
<td>IncomeDeclared</td>
<td>4.587664</td>
<td>1</td>
<td>2.141883</td>
</tr>
<tr>
<td>CreditHistoryLengthMonths</td>
<td>2.168201</td>
<td>1</td>
<td>1.472481</td>
</tr>
<tr>
<td>DebtPaymentsMonthly</td>
<td>1.187513</td>
<td>1</td>
<td>1.089731</td>
</tr>
<tr>
<td>RevolvingCreditAuthorized</td>
<td>3.296183</td>
<td>1</td>
<td>1.815539</td>
</tr>
<tr>
<td>RevolvingCreditUsed</td>
<td>2.437097</td>
<td>1</td>
<td>1.561120</td>
</tr>
<tr>
<td>MaximumCreditLimit</td>
<td>1.378579</td>
<td>1</td>
<td>1.174129</td>
</tr>
<tr>
<td>PaymentsDeclared</td>
<td>6.655463</td>
<td>1</td>
<td>2.579818</td>
</tr>
<tr>
<td>IsInCreditBureau</td>
<td>2.260417</td>
<td>1</td>
<td>1.503468</td>
</tr>
<tr>
<td>IncomeDebtRatio</td>
<td>4.402648</td>
<td>1</td>
<td>2.098249</td>
</tr>
<tr>
<td>MonthDue</td>
<td>1.356867</td>
<td>6</td>
<td>1.025758</td>
</tr>
<tr>
<td>SexM</td>
<td>1.111376</td>
<td>1</td>
<td>1.054218</td>
</tr>
<tr>
<td>catEducationMedia</td>
<td>2.998067</td>
<td>1</td>
<td>1.731493</td>
</tr>
<tr>
<td>catEducationSuperior</td>
<td>3.481766</td>
<td>1</td>
<td>1.865949</td>
</tr>
<tr>
<td>catMaritalStatusCouples</td>
<td>3.259920</td>
<td>1</td>
<td>1.805525</td>
</tr>
<tr>
<td>catMaritalStatusSingle</td>
<td>3.468461</td>
<td>1</td>
<td>1.862380</td>
</tr>
<tr>
<td>catHomeOwnershipFamily</td>
<td>2.078638</td>
<td>1</td>
<td>1.441748</td>
</tr>
<tr>
<td>catHomeOwnershipOwn</td>
<td>2.068247</td>
<td>1</td>
<td>1.438140</td>
</tr>
<tr>
<td>catEmploymentIndependent</td>
<td>1.677869</td>
<td>1</td>
<td>1.295326</td>
</tr>
<tr>
<td>catEmploymentLowValue</td>
<td>1.628873</td>
<td>1</td>
<td>1.276273</td>
</tr>
</tbody>
</table>

* Dependent variable used: BIN_FIRST_RA
Table 3.4. Correlation Matrix: LEVELS
<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>Dependents</th>
<th>MonthsInCurrentResidence</th>
<th>MonthsInCurrentEmployment</th>
<th>IncomeDeclared</th>
<th>CreditHistoryLengthMonths</th>
<th>DebtPaymentsMonthly</th>
<th>RevolvingCreditAuthor</th>
<th>RevolvingCreditUsed</th>
<th>MaximumCreditLimit</th>
<th>PaymentsDeclared</th>
<th>IncomeDebtRatio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>1</td>
<td>0.094673875</td>
<td>0.138044423</td>
<td>0.413002385</td>
<td>0.19657714</td>
<td>0.176861167</td>
<td>0.079334847</td>
<td>0.053462846</td>
<td>0.037058279</td>
<td>0.029701958</td>
<td>0.138044423</td>
<td>0.094673875</td>
</tr>
<tr>
<td>Dependents</td>
<td>0.094673875</td>
<td>1</td>
<td>-0.028041596</td>
<td>0.034886729</td>
<td>0.016542978</td>
<td>0.04978679</td>
<td>0.010842503</td>
<td>0.01089655</td>
<td>-0.00533979</td>
<td>0.020239897</td>
<td>-0.028041596</td>
<td>-0.028041596</td>
</tr>
<tr>
<td>MonthsInCurrentResidence</td>
<td>0.138044423</td>
<td>-0.028041596</td>
<td>1</td>
<td>0.112695158</td>
<td>-0.053224158</td>
<td>0.129723647</td>
<td>-0.028898996</td>
<td>0.064773784</td>
<td>0.032379639</td>
<td>0.02191096</td>
<td>0.112695158</td>
<td>0.112695158</td>
</tr>
<tr>
<td>MonthsInCurrentEmployment</td>
<td>0.413002385</td>
<td>0.034886729</td>
<td>0.112695158</td>
<td>1</td>
<td>0.12425747</td>
<td>0.108165126</td>
<td>0.003141765</td>
<td>0.040631438</td>
<td>0.017309078</td>
<td>0.011630918</td>
<td>0.012695158</td>
<td>0.02191096</td>
</tr>
<tr>
<td>IncomeDeclared</td>
<td>0.19657714</td>
<td>0.016342978</td>
<td>-0.053224158</td>
<td>0.12425747</td>
<td>1</td>
<td>0.008714021</td>
<td>-0.01076509</td>
<td>0.041238325</td>
<td>0.059929326</td>
<td>0.02176627</td>
<td>0.12425747</td>
<td>0.12425747</td>
</tr>
<tr>
<td>CreditHistoryLengthMonths</td>
<td>0.176861167</td>
<td>0.04978679</td>
<td>0.129723647</td>
<td>0.108165126</td>
<td>0.008141021</td>
<td>1</td>
<td>0.157407056</td>
<td>0.350198507</td>
<td>0.226654347</td>
<td>0.146354013</td>
<td>0.176861167</td>
<td>0.176861167</td>
</tr>
<tr>
<td>DebtPaymentsMonthly</td>
<td>0.079334847</td>
<td>0.018842503</td>
<td>-0.028898996</td>
<td>0.03141765</td>
<td>-0.01076509</td>
<td>0.157407056</td>
<td>1</td>
<td>0.38375827</td>
<td>0.480976323</td>
<td>0.424025328</td>
<td>0.079334847</td>
<td>0.079334847</td>
</tr>
<tr>
<td>RevolvingCreditAuthorized</td>
<td>0.053462846</td>
<td>0.01089653</td>
<td>0.064773784</td>
<td>0.040631438</td>
<td>0.04138325</td>
<td>0.350198507</td>
<td>0.38375827</td>
<td>1</td>
<td>0.785803464</td>
<td>0.473087747</td>
<td>0.053462846</td>
<td>0.053462846</td>
</tr>
<tr>
<td>RevolvingCreditUsed</td>
<td>0.037058279</td>
<td>-0.05339739</td>
<td>0.032379639</td>
<td>0.017309078</td>
<td>0.059929326</td>
<td>0.226654347</td>
<td>0.480976323</td>
<td>1</td>
<td>0.396191028</td>
<td>0.045965628</td>
<td>0.037058279</td>
<td>0.037058279</td>
</tr>
<tr>
<td>MaximumCreditLimit</td>
<td>0.029701958</td>
<td>0.020239897</td>
<td>0.02191096</td>
<td>0.011630918</td>
<td>0.02176627</td>
<td>0.146354013</td>
<td>0.424025328</td>
<td>0.473087747</td>
<td>0.396191028</td>
<td>-0.018851038</td>
<td>0.037058279</td>
<td>0.037058279</td>
</tr>
<tr>
<td>PaymentsDeclared</td>
<td>0.146873777</td>
<td>0.107270674</td>
<td>-0.123817112</td>
<td>0.055718611</td>
<td>0.637438813</td>
<td>0.01174244</td>
<td>-0.036376765</td>
<td>0.02979196</td>
<td>0.045965628</td>
<td>-0.018851038</td>
<td>0.146873777</td>
<td>0.146873777</td>
</tr>
<tr>
<td>IncomeDebtRatio</td>
<td>0.032347081</td>
<td>0.102060386</td>
<td>-0.069404824</td>
<td>-0.040978339</td>
<td>-0.16032236</td>
<td>8.066-06</td>
<td>-0.028153763</td>
<td>0.11331998</td>
<td>0.009787704</td>
<td>-0.04065321</td>
<td>0.032347081</td>
<td>0.032347081</td>
</tr>
</tbody>
</table>
3.4. Regressions, CARTS, Random Forests, and “Clustering-then-Predict” approach.
In the following section, we present the results for the three main methodologies used in our research: Logistic Regressions, Classification and Regression Trees, and Random Forests. It is important to mention that our results are divided in two main approaches including: 1) traditional direct prediction over whole sample data; and 2) a Clustering-then-predict technique.

As is widely known in machine learning and other data analytics applications, accuracy tends to improve when customized regressions are run over specific clusters (vis-a-vis models trained over the whole sample). In our case, the clustering technique used was *k*-means clustering as it lends itself better for prediction purposes. We present the algorithm used below:

<table>
<thead>
<tr>
<th>K-MEANS CLUSTER ALGORITHM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Specify desired number of clusters (we used Hierarchical clustering and corresponding Dendrogram to derive the number of clusters)</td>
</tr>
<tr>
<td>2. Randomly assign each data point to a cluster</td>
</tr>
<tr>
<td>3. Calculate cluster centroids</td>
</tr>
<tr>
<td>4. Re-assign each data point to closest cluster centroid</td>
</tr>
<tr>
<td>5. Repeat steps 4) and 5) until no further improvement is made</td>
</tr>
</tbody>
</table>

It is noteworthy to mention that to apply the k-means clustering process we needed to normalize our data set as part of the centroid calculation in step 2. Moreover, as mentioned in step 1, we previously run a Hierarchical Clustering Algorithm to generate a Dendrogram where we specified 2 as the number of clusters to be used in our analysis (See Figure 3.2 below)

**Figure 3.2. Hierarchical Clustering Dendrogram**
Once we specified To do this, we followed a procedure similar to the transformation of categorical levels into binary variables discussed above, the only difference is that for clustering purposes we did not grouped the levels into larger categories (e.g. Medium Education (= High School + Technical), and generated a binary variable per each level within each categorical to get a finer detail in our cluster generation (i.e. Education (= 6 binaries); Marital Status (= 6 binaries); Homeownership (4 binaries); Employment (26 binaries)). Once we generated our k-means clusters, the number of observation per cluster was as follows:

- **K-means Cluster 1**: 1007 observations
- **K-means Cluster 2**: 993 observations

Employing this clustering, our observations were further subdivided in training and testing sets in the following way:
• **Training set corresponding to Cluster 1** = 704 observations

• **Training set corresponding to Cluster 2** = 695 observations

• **Testing set corresponding to Cluster 1** = 303 observations

• **Testing set corresponding to Cluster 2** = 298 observations

This sets were utilized to train Logistic, CART, and Random Forest models as will be shown below in Section 3.6.

### 3.4.1. Direct approach: Model’s Predictions over whole data set

In the following section we present the results from the Logistic Regressions, Classification Trees and Random Forest models we applied over our complete sample (i.e. without clustering). As explained in Section 2.4 we divided our data set between *Training* and *Testing* subsets to reduce overfitting effects over our model’s results. The split ratio utilized was: training set (70%); testing set (30%).

It is noteworthy to mention that we included a third set for validation purposes at the following ratio: *training* (60%); *validation* (20%); and *testing* (20%); however, the set size reduction more than offset the benefits of including an intermediate validation step given the limited amount of observations in our sample. The final decision, thus, was to run our results for the two sets and split rate mentioned above.

We run a total of 40 models, applying the Logit, CART, and Random Forest methodologies over 8 versions of our dependent variable (see Table 3.1. Variable Dictionary above). Our models were run both on simple binary versions of the dependent, as well as in a 5-class version to
detect more granular results in the default behavior. Moreover, we run all our models with and without the logarithmic transformations mentioned above, consistently getting better results in the latter case. Table 3.2 presents a heat-map with the summary of results for this regressions. In the coming sections we present the best results for each methodology.

**Direct approach: Logistic Regressions**

For our Logistic Regressions we run a semi-automatic *Stepwise Regression* procedure over a set of 14 numerical and 11 factor variables (See Appendix A1.2.). The basic idea of the procedure is to depart from a given model - in our case our initial model started regressing all the independents (See results in Table 3.6 below)-, and then adding or deleting independent variables until reaching that model with the best results under AIC criteria. This eliminates multicollinearity issues, while identifying those variables with higher explanatory power over default probability. For the original model containing all independent variables we have a testing accuracy ratio \[= (TN+TP) / (TN+FN+TP+FP)\] of 68\%34.

---

34 Training accuracy rate for this model containing all independent variables was 67.43\%.
Table 3.5. Summary of results for all models (Direct and “Clustering-then-Predict” approaches)

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>BinFull_RA</th>
<th>BinFull_RP</th>
<th>ClassPF_RA</th>
<th>ClassPF_RP</th>
<th>BinFP_RA</th>
<th>BinFP_RP</th>
<th>BinFirstRA</th>
<th>BinFirstRP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logit stepwise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accu.Train Set</td>
<td>61.67%</td>
<td>64.33%</td>
<td>63.72%</td>
<td>48.63%</td>
<td>63.89%</td>
<td>63.00%</td>
<td>64.28%</td>
<td>67.15%</td>
</tr>
<tr>
<td>Accu.Test Set</td>
<td>61.00%</td>
<td>63.67%</td>
<td>62.90%</td>
<td>47.83%</td>
<td>64.17%</td>
<td>63.17%</td>
<td>65.67%</td>
<td>67.50%</td>
</tr>
<tr>
<td>Logit tuned</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accu.Test Set</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>66.73%</td>
<td>68.50%</td>
</tr>
<tr>
<td>CART</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accu.Train Set</td>
<td>67.07%</td>
<td>66.92%</td>
<td>65.26%</td>
<td>52.14%</td>
<td>65.14%</td>
<td>65.50%</td>
<td>69.33%</td>
<td>71.42%</td>
</tr>
<tr>
<td>Accu.Test Set</td>
<td>58.33%</td>
<td>61.50%</td>
<td>60.73%</td>
<td>49.83%</td>
<td>64.16%</td>
<td>62.33%</td>
<td>64.17%</td>
<td>67.00%</td>
</tr>
<tr>
<td>Random Forest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accu.Test Set</td>
<td>60.29%</td>
<td>62.93%</td>
<td>63.90%</td>
<td>47.71%</td>
<td>62.07%</td>
<td>63.85%</td>
<td>65.07%</td>
<td>71.21%</td>
</tr>
<tr>
<td>CART cp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accu.Train Set</td>
<td>63.83%</td>
<td>61.00%</td>
<td>63.06%</td>
<td>48.50%</td>
<td>64.67%</td>
<td>62.67%</td>
<td>64.00%</td>
<td>66.67%</td>
</tr>
<tr>
<td>Accu.Test Set</td>
<td>59.80%</td>
<td>63.00%</td>
<td>62.90%</td>
<td>47.67%</td>
<td>63.00%</td>
<td>63.17%</td>
<td>64.33%</td>
<td>57.20%</td>
</tr>
<tr>
<td>RF cp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accu.Train Set</td>
<td>64.00%</td>
<td>63.15%</td>
<td>61.96%</td>
<td>48.00%</td>
<td>63.00%</td>
<td>65.50%</td>
<td>65.00%</td>
<td>67.80%</td>
</tr>
<tr>
<td>Accu.ALL Train</td>
<td>60.50%</td>
<td>61.60%</td>
<td>63.11%</td>
<td>49.58%</td>
<td>63.60%</td>
<td>63.95%</td>
<td>65.82%</td>
<td>68.43%</td>
</tr>
<tr>
<td>Accu.Test 1</td>
<td>60.93%</td>
<td>62.58%</td>
<td>64.45%</td>
<td>49.67%</td>
<td>63.35%</td>
<td>64.00%</td>
<td>65.23%</td>
<td>68.20%</td>
</tr>
<tr>
<td>Accu.Test 2</td>
<td>61.74%</td>
<td>62.63%</td>
<td>64.43%</td>
<td>49.50%</td>
<td>64.28%</td>
<td>62.72%</td>
<td>66.77%</td>
<td>67.41%</td>
</tr>
<tr>
<td>CART cp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accu.Train Set</td>
<td>63.33%</td>
<td>62.60%</td>
<td>62.60%</td>
<td>49.58%</td>
<td>63.50%</td>
<td>63.83%</td>
<td>65.83%</td>
<td>68.05%</td>
</tr>
<tr>
<td>Accu.Test 1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>66.72%</td>
<td>69.19%</td>
</tr>
<tr>
<td>Accu.Test 2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>67.67%</td>
<td>68.52%</td>
</tr>
<tr>
<td>Accu.ALL Train</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>67.32%</td>
<td>68.90%</td>
</tr>
<tr>
<td>CART</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accu.Train 1</td>
<td>72.20%</td>
<td>75.46%</td>
<td>70.82%</td>
<td>60.00%</td>
<td>74.60%</td>
<td>75.15%</td>
<td>81.66%</td>
<td>76.61%</td>
</tr>
<tr>
<td>Accu.Train 2</td>
<td>73.96%</td>
<td>68.10%</td>
<td>68.35%</td>
<td>58.93%</td>
<td>74.71%</td>
<td>76.07%</td>
<td>74.52%</td>
<td>74.71%</td>
</tr>
<tr>
<td>Accu.ALL Train</td>
<td>73.71%</td>
<td>71.81%</td>
<td>69.95%</td>
<td>59.47%</td>
<td>74.66%</td>
<td>75.64%</td>
<td>76.21%</td>
<td>74.71%</td>
</tr>
<tr>
<td>CART cp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accu.Train 1</td>
<td>65.66%</td>
<td>64.25%</td>
<td>62.18%</td>
<td>48.65%</td>
<td>64.53%</td>
<td>63.19%</td>
<td>63.83%</td>
<td>66.46%</td>
</tr>
<tr>
<td>Accu.Train 2</td>
<td>60.57%</td>
<td>59.62%</td>
<td>60.39%</td>
<td>46.54%</td>
<td>59.91%</td>
<td>62.63%</td>
<td>63.16%</td>
<td>67.20%</td>
</tr>
<tr>
<td>Random Forest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accu.Train 1</td>
<td>60.57%</td>
<td>61.96%</td>
<td>61.24%</td>
<td>47.61%</td>
<td>62.47%</td>
<td>62.42%</td>
<td>63.50%</td>
<td>66.86%</td>
</tr>
<tr>
<td>Accu.Train 2</td>
<td>63.90%</td>
<td>64.57%</td>
<td>66.11%</td>
<td>50.00%</td>
<td>63.24%</td>
<td>63.34%</td>
<td>64.57%</td>
<td>66.67%</td>
</tr>
<tr>
<td>Accu.ALL Test</td>
<td>62.17%</td>
<td>63.94%</td>
<td>64.44%</td>
<td>47.59%</td>
<td>63.61%</td>
<td>63.33%</td>
<td>64.39%</td>
<td>67.55%</td>
</tr>
<tr>
<td>CART cp</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>64.56%</td>
<td>61.92%</td>
</tr>
<tr>
<td>Accu.Test 1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>61.27%</td>
<td>64.26%</td>
</tr>
<tr>
<td>Accu.Test 2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>64.24%</td>
<td>63.34%</td>
</tr>
<tr>
<td>Random Forest cp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accu.Train 1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>64.24%</td>
<td>63.34%</td>
</tr>
<tr>
<td>Accu.Train 2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>69.94%</td>
<td>63.94%</td>
</tr>
<tr>
<td>Accu.ALL Test</td>
<td>61.83%</td>
<td>62.94%</td>
<td>64.27%</td>
<td>48.42%</td>
<td>62.60%</td>
<td>63.67%</td>
<td>65.16%</td>
<td>68.22%</td>
</tr>
</tbody>
</table>
Table 3.6. Logistic Regression obtained through AIC step-wise procedure

| Coefficients          | Estimate | Std. Error | z value | Pr(>|z|) |
|-----------------------|----------|------------|---------|---------|
| (Intercept)           | -0.17868 | 1.24837    | -0.143  | 0.88619 |
| Age                   | -0.04540 | 0.00794    | -5.734  | 3.1e-08 *** |
| Dependents            | 0.074609 | 0.033998   | 2.218   | 0.02667 |
| MonthsInCurrentResidence | 0.00197 | 0.001321   | 0.946   | 0.34502 |
| IncomeDeclared        | 0.181288 | 0.131714   | 1.392   | 0.16406 |
| DebtPaymentsMonthly   | 0.008408 | 0.020000   | 0.420   | 0.67417 |
| RevolvingCreditUsed   | 0.023828 | 0.021363   | 1.126   | 0.26019 |
| MaximumCreditLimit    | -0.005332 | 0.013758   | -0.394  | 0.69439 |
| IsInCreditBureau      | 0.074192 | 0.130426   | 0.570   | 0.56842 |
| MonthDue2             | -0.152534 | 0.178707   | -0.857  | 0.39181 |
| MonthDue3             | 0.267740 | 0.187883   | 1.417   | 0.15763 |
| MonthDue4             | -0.049436 | 0.369239   | -0.134  | 0.89349 |
| MonthDue5             | -0.279210 | 0.374214   | -0.746  | 0.45559 |
| MonthDue6             | 0.247436 | 0.370824   | -0.667  | 0.49662 |
| MonthDue7             | 0.013168 | 0.389623   | 0.034   | 0.97304 |
| SexM                  | 0.075671 | 0.127826   | 0.615   | 0.53882 |
| cEducationMedia       | -0.057920 | 0.202571   | -0.286  | 0.77494 |
| cEducationSuperior    | -0.553761 | 0.220418   | -2.512  | 0.01199 ** |
| catMaritalstatusCouples | -0.609668 | 0.210388   | -2.905  | 0.003766 *** |
| catMaritalstatussingle | -0.088784 | 0.223732   | -0.394  | 0.69439 |
| cHomeOwnershipFamily  | 0.099110 | 0.167302   | 0.592   | 0.55358 |
| cHomeOwnershipPnon    | 0.062973 | 0.170434   | 0.372   | 0.70961 |
| cEmploymentindependent| 0.190219 | 0.152554   | 1.247   | 0.21243 |
| cEmploymentLowvalue   | 0.018136 | 0.164486   | 0.110   | 0.91220 |
| DebtIncomeRatio       | -0.044116 | 0.108644   | -0.413  | 0.67968 |

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1784.1 on 1399 degrees of freedom
Residual deviance: 1705.6 on 1374 degrees of freedom
AIC: 1757.6

Number of Fisher Scoring iterations: 4

After running the stepwise for all our versions of the dependent variable, we detected that BIN_FIRST_RP (i.e. binary version based on number of days of delayed payment relative to the original loan due date based on a Risk Averse lending criteria) provided the best results both in terms of accuracy rate in a testing set with 67.50% and AUC (area under ROC curve) with a 62.12% statistic (see Appendix 3.3). These results are superior by 3.5 percent over a basic baseline’s results.35

After obtaining this result, we evaluated different versions of the regression looking for an improvement in the prediction while making sense from an economical point of view. The result

---

35 The baseline is defined as: the number of most common outcome in the sample / total number of observations ≈ 64%
is shown in Table 3.7 providing an accuracy rate of 68.50% and an AUC statistic of 62.19%. This regression is the second highest result among all the 45 models and versions of the dependent variables we tested (in out-of-sample sets).

The next best results were obtained regressing over BINPF_RA (the version of the dependent including both the initial and extended-period default behavior, see dictionary above), having an accuracy rate of 66.73% and an AUC statistic of 61.38%.

As we can see from Table 3.7, the probability of default (PD) is inversely related with: Age, high level of education, and single/married marital status (as opposed to divorced, separated, or widowed borrowers). All of these variables are statistically significant at least at a 95% level (3 are significant at a 99% and higher). On the other hand, PD seems to be positively related with: the number of borrower’s dependents, and level of leverage (as expressed by debt-to-income-ratio), although, both variables are not statistically significant.

Table 3.7. Logistic Regression obtained through AIC step-wise (improved version)

```r
> summary(LogModel)

Call: 
glm(formula = Dependent ~ Age + Dependents + DebtIncomeRatio + 
catEducationSuperior + catMaritalStatusCouples + catMaritalStatusSing 
le, family = binomial, data = Train)

Deviance Residuals:
  Min 1Q Median 3Q Max
-1.4449 -0.9173 -0.7658 1.2789 2.1006

(Dispersion parameter for binomial family taken to be 1)

    Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.5050249 0.3552463 4.237 2.27e-05 ***
Age -0.0443046 0.0069910 -6.337 2.34e-10 ***
Dependents 0.0754038 0.0523617 1.440 0.14985
DebtIncomeRatio 0.0004005 0.0784094 0.005 0.99593
catEducationSuperior -0.3607784 0.1199827 3.007 0.00264 **
catMaritalStatusCouples -0.5961482 0.2066365 -2.885 0.00391 **
catMaritalStatusSingle -0.5153128 0.2214396 -2.327 0.01996 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

102
Nevertheless, the inclusion of these variables contribute in increasing the predictability of the model and provides an interesting causality from an economic point of view, i.e. that defaulting behavior tends to increase as the person’s overall level of economic commitments augments, as implied by a larger number of dependents (family responsibilities) and previous debts.

Based on these results, we can infer with almost more than 95% confidence that a senior borrower with a bachelor (or higher levels of education), who is whether single or married will in average will tend to default less than a younger individual, with only basic or medium level of education, widowed or separated, who is in charge of a large household. The question now is what are the precise age, debt leverage, or number of dependents thresholds distinguishing between a failing and non-failing loans. As we will see below, this is where other methodologies such as CART can complement our Logistic findings by providing detail on the size of these decision splits.

**Figure 3.3. ROC for model based on BIN_FIRST_RA dependent variable**
As mentioned in chapter 2, this detailed breakdown of results can be utilized to support additional areas in the lending process. For instance, the lender can redefine his market targeting strategy focusing on individuals with the above cited characteristics (if and only if his main objective is just the minimization of the default event). For a detailed comparison of the results for the remaining dependent variable versions see Table 3.2.

Before continuing it is important to mention that during our Logistic regressions (and CART models) we observed that when including the *Income Declared* variable, it yielded a positive relationship with our dependent(s) variable(s). This relation seem counterintuitive given that a person with higher declared income would seem to display a higher probability of default.

From some preliminary results in CART we detected that this relation is not direct but appears to be closely related to a person’s overall indebtedness. That is, although a person might have a higher level of income, a large level of past debts may more than impair his payment capacity, *ergo* such individual could display a higher PD versus a person with less income but who also is less financially leveraged.

A second explanation to this strange sign could be that some obligors might be overstating their level of income to ensure the loan approval. As we mentioned above, Mimoni lending process is based on an online application that does not require physical proof of income given that this information is cross-validated with credit bureau information as part of their algorithm. Therefore, this “flexibility” of requirements might be leading to a consistent overstatement of incomes in individuals who tend to default more often. That would explain the strange sign in regressions where we include other significant predictors but not in univariate models (i.e. Dependent vs. Income Declared).
The inclusion of additional independent variables might just expose the inconsistency of the stated income relative to the other variables’ levels (see Appendix A1.4. for an example of progressive regressions to check this sign shift in the income coefficient). If this hypothesis is true, the lender could detect “inconsistent” applications using inflated incomes (relative to their other variables) as a significant predictor of default. As mentioned above, CART might be used to determine appropriate thresholds between inflated and normal applications.

This results seem plausible given that we already controlled for multicollinearity eliminating any variables with correlation coefficients superior to 0.61, and VIFs above 5 (See Tables 3.3, 3.4 above and A1.1); and, these procedures did not solve the “strange” sign problem (this might support the abovementioned hypothesis). This might imply that the behavior is representative and is not due to multicollinearity distortions, but captures a true behavior in the borrower sample (see Direct CART section below for further discussion).

Once controlling for multicollinearity, we also inspected the pairwise correlation with the dependent variable to see if the relation was due to our specific sample. The results are shown in Table 3.7 and as we can appreciate the relationship displays a *logical* negative sign, having higher default behavior in obligors with lower levels of declared income and vice versa.

**Table 3.7. Correlation Matrix: Dependent, Income Declared**

<table>
<thead>
<tr>
<th></th>
<th>Dependent</th>
<th>IncomeDeclared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent</td>
<td>1</td>
<td>-0.037993698</td>
</tr>
<tr>
<td>IncomeDeclared</td>
<td>-0.037993698</td>
<td>1</td>
</tr>
</tbody>
</table>
Direct approach: Classification and Regression Trees (CART)

As mentioned above, among some of the advantages of CART models over Logistic regressions are: 1) that it does not assume (and thus is not limited to) linear relationships in the data; and 2) is more interpretable given the visual splitting and decision rules derived from its recursive partitioning.

It is important to mention that we run our trees both using the default splitting parameters, as well as a cross-validation pruning technique defining the optimal number of folds (c.p.) for our data based on an accuracy criteria. By setting this limit on the number of observations in each subset, we control the level of complexity of our trees, that is, the smaller the c.p. parameter the more splits (ergo more complex) our tree will be. Thus if c.p. is too small, overfitting might occur, while if it is too large, the model will be too simple and accuracy will be poor. The range of c.p.'s used in our trees was [0.014, 0.1].

Once again, the best results among all our dependent variable versions were for the BIN_FIRST_RP with and accuracy rates of 71.42% in the training set and 67% for out-of-sample testing, the AUC for the testing set was 61.20%. As in our Logistic findings, the second best results were obtained using BIN_FIRST_RA with accuracy results of 69.35% and 64.17% for the training and testing sets respectively.

On other hand, it is important to mention that the cross-validation results (using optimal c.p. parameter of 0.014) present a slight improvement in the testing set reaching an out-of-sample accuracy of 64.67%. Figure 3.4 presents the CART model obtained using default c.p. parameters
Our CART results seem to be consistent and complementary with those obtained through Logistic regression. As mentioned, one of the advantages of trees is that they provide further detail on the specific splits between failing and non-failing loans. It is noteworthy to mention that in all our CARTS the first split starts with the age of the borrower, differentiating somehow between young or senior obligors. It is particularly interesting that the threshold age is around the early 30s, implying that borrowers below that age (provided his specific levels in other variables) tend to default more than more senior obligors.

Additionally, from these results we can corroborate that the number of dependents is positively related with higher probabilities of default, while a higher level of education (see A1.5) implies that people will tend to pay their commitments more often. Additionally, as expected, higher levels of indebtedness (as expressed in a large Debt-Income Ratio) and an exercise use of a borrower’s authorized credit (i.e. Revolving Credit Used) seem to be more related with higher defaulting probabilities.

On other hand, an interesting point that cannot be fully appreciated in the Logistic results are the dynamics behind a high level of declared income and default behavior. As we saw in our step-wise regressions, a higher level of declared income seems to be related with a higher probability of default, a fact that seems counterintuitive (i.e. higher declared income normally is associated with stronger capacity to meet one’s obligations).

As mentioned above, Mimoni’s online process does not require borrowers to provide proof-of-income in its application, given that one of its competitive advantages is providing a quick and
hassle-free credit application experience. Mimoni algorithm instead cross-validates this information with credit bureau reports and proprietary socio-economic profiles and established whether the candidate is subject of credit. However, this requirement-free experience might be triggering an income overstating behavior in a particular segment of applicants (that happened to have a higher default behavior).

As we can see in Table 3.4 and in CARTs in A.1.4, a normal relationship between Income Declared and the Dependent variable holds as long as income stay around a 8.3 threshold (= log(Income Declared + 1)), but the distortion seems to appear when a borrower declares incomes above 9 but his other variables are more representative of lower income levels (i.e. overstatement, artificial income inflation). For additional examples on this dynamics see CARTs in A1.5, A1.6.

As mentioned above, if this hypothesis is true, the lender could detect “inconsistent” applications using this overstated variable as a significant predictor of default given that consistently this inflated incomes are highly related with large PDs. In further research, this findings could be complemented with other data mining methods to detect more precisely when a specific application presents inconsistencies relative to the average profile (at those variable levels).

As we already mentioned above, one of the strongest advantages of CART models is its interpretability and capacity to be used as a guide for informed lending decision processes. This sort of “causality” can hardly be discerned directly from our Logistic regression coefficients alone, thus, CART (and other data mining solutions) can be used to complement and guide loan managers towards more precise and informed decisions.
Direct approach: Random Forests (RF)

As mentioned in our literature review, Random Forest is another type of recursive partitioning method involving a group of classification trees (i.e. ensemble learning method) that are
computed on random subsets of a sample utilizing a group of random predictors for each split in each CART. The results of Random Forests have been demonstrated to generate better and more accurate predictions than the results of standalone CARTs.

However, it is important to mention that given that RFs are built through the generation of a large collection of trees, we lose some of the interpretability that comes from single CARTs in terms of visually understanding how predictions are made (interrelation) and which variables are more important. Nevertheless, we will include two techniques that will allow us to identify the contribution and behavior of the most important predictors across the forests: 1) the number of times that a variable is selected for a split along all trees in the forest, and 2) an impurity measure which indicates how homogeneous each leaf in a tree is (i.e. reduced impurity higher importance of a variable). Finally, as in the case of CARTs, we run our Random Forests using both default and accuracy-optimized c.p. parameters.

In the case of RFs, the best results based on optimal c.p. parameters were obtained taking BIN_FIRST_RP as our dependent variable. The accuracy rates were of 67.20% for training set and 67.80% for out-of-sample data. It is noteworthy to mention that, as expected, RFs in general generated better accuracy rates than standalone CARTS using both default and optimized cross-validation parameters. On other hand, the second best out-of-sample results were for BIN_FP_RA reaching a 64.67% accuracy under default parameters, and BIN_FP_RP with 65.50% applying optimized c.p.

As we mentioned above, although RF do not have the visual advantage of portraying variable relationships into a tree (by definition RF is made of hundreds of trees), we can still extract valuable information about: 1) which predictors are more often utilized along the trees in the
forest, hence are more important as discriminatory variables; and 2) the quality and purity of the splits utilized. In the coming lines, we will explore in more detail this information.

In Figure 3.5.A we can observe which variables were more often used in the partitioning process and the number of times it was utilized (see Appendix A1.5 for the algorithm used to generate this table in R).

Additionally, on Figure 3.5.B we introduced our model’s GINI Impurity chart, which measures the homogeneity in each bucket. GINI impurity is an indicator of the times that a randomly selected observation would be incorrectly labeled in a wrong bucket. Under this definition, every time we select a variable to perform a split, impurity will decrease. This implies that we can estimate the relative importance of a variable by measuring the reduction in impurity that results from selecting a variable for splitting across all trees in a RF.

From Figure 3.5.A and B, we see that our RF results are consistent with CART and Logistic regression. In both diagrams, we can clearly observe that as before Age has a preponderant importance in the partitioning process, being the variable with the largest number of partitions across the trees (more than 16,000) and better GINI coefficient (above 80).

Complementing this finding with our CARTs and Logit models, therefore, it is evident that Age should be taken as a pivotal variable in the credit concession process. Moreover, from our CART results we know that there is a specific threshold, around their early 30s, where obligors start displaying distinct defaulting pattern.
A remarkable point from these diagrams is that there are eight independent variables that stand out from the rest both in terms of number of splits (12,000, 16,000) and bucket purity, these variables are: Age; Months in current employment and residence (as measures of physical stability); Income Declared (with the overstating qualification mentioned above); Debt Payments Monthly and Debt Income Ratio (as measures of financial leverage); and Maximum Credit Limit. It is significant to highlight that RF allows us to identify variables whose importance was not fully appreciated in Logistic regressions and CART. Specifically, work and residence stability seem to be particularly important as discriminatory variables between defaulters and non-defaulters.

Once again, we can see that the use of combined methods allows us to grasp a broader and finer picture of the dynamics behind the defaulting phenomena that would be otherwise inaccessible through the use of isolated methods as is the widespread practice in the low-income consumer lending industry. This is one of the drivers of the present research.

Figure 3.5.A. Diagram with number of times a variable is used for a split in RFs
Figure 3.5.B. Impurity chart for RF model run on BIN_FIRST_RP
3.4.2. “Clustering-then-Predict” (CTP) approach: tailoring regressions to specific groups

In the following section we present our results from the Logistic Regressions, Classification Trees and Random Forest models preforming a series of clustering procedures and then fitting cluster-specific prediction models for each group. Among some of the main advantages of this methodology we have: 1) the development of more group specific models tends to increase the accuracy power of our predictions, and 2) clusters are interpretable and can reveal unique patterns that can be highly useful in the formulation of group-specific strategies.

As explained in the beginning of these section, our first step was to determine the number of clusters using the hierarchical-clustering dendrogram shown in Figure 3.2. It is important two mention that during our regressions we run our models using both 2 and 3 clusters, obtaining consistently the best results under the 2-groups option\(^{36}\). The actual cluster split was conducted following a k-means algorithm obtaining a well-balanced number of observations between the 2 clusters. After clustering, we applied a 70/30 split on each cluster obtaining on average the expected 700 observations on training and 300 observations on testing sets (see above for number of elements in each cluster and in each training/testing set).

Some interesting insights about the composition of each cluster can be appreciated from Table 3.8. For instance, we can see that in average borrowers in cluster 1 are 4.1 years older than those in cluster 2; have a slightly larger household (1 more dependent); earn approximately $4,500MX more (as declared in their application); have almost twice as much debts registered in the credit bureau; and are 0.3 times more indebted than their cluster 2 counterparts (see below for further

\(^{36}\) This could be the result of the limited number of observations to which we have access and specially its impact in testing-set size (i.e. after dividing in 3 clusters we obtained an average of 200 per set). Nevertheless, it is highly recommended to conduct further research with more granular clustering, because the 3-group option gave results close to our 2-cluster findings despite the reduced number of observations.
comparison). Additionally, given the relative proportions in the Marital Status category, it appears that Cluster 1 has more married people than Cluster 2, which might explain many of the dynamics between PD and the statistically significant predicting variables in this group.

Table 3.8. Mean comparison between members of Cluster 1 and Cluster 2

<table>
<thead>
<tr>
<th>MEAN</th>
<th>Age</th>
<th>Dependent</th>
<th>MonthsInCurrentResidence</th>
<th>MonthsInCurrentEmployment</th>
<th>IncomeDeclared</th>
<th>CreditHistoryLengthInMonths</th>
<th>DebtPaymentsMonth</th>
<th>RevolvingChargeRatio</th>
<th>MaximumCreditLimit</th>
<th>PaymentsDeclared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>39.18</td>
<td>1.597</td>
<td>16.05</td>
<td>7.179</td>
<td>13502</td>
<td>3.721</td>
<td>22581.8</td>
<td>4452</td>
<td>3688</td>
<td>9260.6</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>35.08</td>
<td>0.8127</td>
<td>16.97</td>
<td>6.032</td>
<td>8898</td>
<td>2.985</td>
<td>11692.9</td>
<td>2016</td>
<td>1224</td>
<td>6153</td>
</tr>
<tr>
<td>Difference</td>
<td>4.1</td>
<td>0.7843</td>
<td>-0.92</td>
<td>1.147</td>
<td>4604</td>
<td>0.736</td>
<td>10888.9</td>
<td>2436</td>
<td>2464</td>
<td>3107.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MEAN</th>
<th>lnCreditBureau</th>
<th>Sex</th>
<th>catEducation</th>
<th>catEducationSuperior</th>
<th>catMaritalStuCouple</th>
<th>catMaritalStuSingle</th>
<th>catHomeOwnership</th>
<th>catHomeOwnership</th>
<th>catEmploymentDependent</th>
<th>catEmploymentLowValue</th>
<th>DebtIncomeRatio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>0.5382</td>
<td>0.9722</td>
<td>0.4081</td>
<td>0.4896</td>
<td>0.7557</td>
<td>0.1331</td>
<td>0.3406</td>
<td>0.4052</td>
<td>0.3873</td>
<td>0.3386</td>
<td>1.7366</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>0.5176</td>
<td>0.3092</td>
<td>0.4502</td>
<td>0.4491</td>
<td>0.3182</td>
<td>0.6073</td>
<td>0.4783</td>
<td>0.3374</td>
<td>0.5146</td>
<td>0.2578</td>
<td>1.443</td>
</tr>
<tr>
<td>Difference</td>
<td>0.0206</td>
<td>0.663</td>
<td>-0.0421</td>
<td>0.0405</td>
<td>0.4375</td>
<td>-0.4722</td>
<td>-0.1377</td>
<td>0.0678</td>
<td>-0.1273</td>
<td>0.0808</td>
<td>0.2936</td>
</tr>
</tbody>
</table>

CTP approach: Logistic Regressions

Just as in the case of Logistic regressions applied over the whole data set (i.e. no clustering), in the CTP approach we started by applying a *stepwise regression* to obtain the best model specification under an AIC criteria for each one of our clusters. To achieve this, our first step was to run a “base” Logistic model, for each group, over the complete list of numerical and categorical variables (21 variables: 9 numerical and 12 categorical; see Appendix A.1.2.B), and generated its correspondent *stepwise*. Once we determined this AIC-optimal form, we continued adding and deleting variables until reaching a model not only with higher accuracy (both in training and testing sets) but also with useful interpretability from an operative point of view.

Table 3.9.A and B present the “*improved*” stepwise specifications for the dependent variable version with best results both at cluster and overall-accuracy levels. Once again, BIN_FIRST_RP provided the best results with an *aggregated* accuracy outcome in the stepwise regressions of
68.05%, and cluster-specific accuracies of 68.20% and 67.41% for Cluster 1, and Cluster 2 respectively.

Moreover, after implementing the accuracy and interpretability tuning we discussed, we achieved our highest accuracy result (among all the 108 models including Logistic, CARTS, RF, and cross-validated decision trees) using BIN_FIRST_RP as our dependent variable. For this model (shown in Figure 3.9), our overall accuracy rate was of 68.90% with and AUC of 63.10%, with group-specific predictions of 69.19% for Cluster 1 and 68.52% for Cluster 2.

These results are superior in almost 5% over the basic baseline results (see “Direct approach” section above). As in the case of direct regressions, the next best results were also obtained using BINPF_RA with aggregated-accuracy of 67.32%, ROC-AUC of 61.35%, and cluster-specific accuracies of 66.72% and 67.67% for Cluster 1 and Cluster 2 respectively. The specifications for are shown below.

As can be observed from Table 3.9.A, for those borrowers included in Cluster 1, the probability of default seems to be negatively related to Age, Month Due 5, and single/married status, with a high statistical significance level (above 99% for four of them, and 95% for the seasonal variable). On the other hand, it seems that, for this cluster, the default behavior tends to increase with the number of people depending economically from the borrower.
Table 3.9.A. Logistic Regression obtained through AIC step-wise (Cluster 1)

```r
> summary(step1)
```

```r
Call:
glm(formula = Dependent ~ Age + Dependents + MonthDue + catEducationSuperior + catMaritalStatusCouples + catMaritalStatusSingle + catHomeOwnershipFamily, family = binomial, data = train1A)

Deviance Residuals:
  Min       1Q   Median       3Q      Max
-1.5429   -0.8870  -0.7101   1.2281   2.1037

Coefficients:
                 Estimate Std. Error  z value  Pr(>|z|)
(Intercept)      1.80657    0.66274    2.726   0.006412 **
Age              -0.03935    0.01037   -3.795   0.000148 ***
Dependents       0.19970    0.07263    2.750   0.006567 **
MonthDue2        -0.45626    0.46057   -0.991   0.321866
MonthDue3        0.05966    0.47408    0.126   0.900662
MonthDue4        -0.18555    0.44671   -0.415   0.677874
MonthDue5        -0.92029    0.48737   -1.969   0.048946 *
MonthDue6        -0.42091    0.44975   -0.936   0.349346
MonthDue7        -0.59726    0.47685   -1.253   0.210384
catEducationSuperior -0.26430    0.17348   -1.524   0.127633
catMaritalStatusCouples -0.86174    0.25947   -3.321   0.000096 ***
catMaritalStatusSingle -0.86331    0.33049   -2.612   0.008997 **
catHomeOwnershipFamily -0.24905    0.18267   -1.363   0.172748

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 3.9.B. Logistic Regression obtained through AIC step-wise (Cluster 2)

```r
> summary(step2)
```

```r
Call:
glm(formula = Dependent ~ Age + DebtPaymentsMonthly + IsInCreditBureau + catEducationSuperior + MonthDue + catEmploymentIndependent, family = binomial, data = train2A)

Deviance Residuals:
  Min      1Q   Median      3Q     Max
-1.4179  -0.9347  -0.7623  1.2779  2.1176

Coefficients:
                 Estimate Std. Error  z value  Pr(>|z|)
(Intercept)      -0.38275     0.75309   -0.508   0.6113
Age              -0.03853     0.00939   -4.104  4.06e-05 ***
DebtPaymentsMonthly 0.04748     0.02171    2.187   0.0288 *
IsInCreditBureau  0.37594     0.16571    2.269   0.0233 *
Year              -0.40765     0.17570   -2.320   0.0203 *
MonthDue2         0.70764     0.68437    1.034   0.3011
MonthDue3         0.81720     0.70524    1.159   0.2466
MonthDue4         0.75245     0.67862    1.109   0.2675
MonthDue5         0.57072     0.68686    0.831   0.4060
MonthDue6         0.61922     0.68380    0.906   0.3652
MonthDue7         0.50036     0.73799    0.678   0.4978
catEmploymentIndependent 0.39900     0.16786    2.358   0.0180 *

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
This last finding seem to be consistent with the attributes described in Table 3.8 where we detected that borrowers in Cluster 1 have in average one more direct dependent in their households. It is important to mention that at these level of income (approximately $13,500MX), the economic expenses related to an additional family member could represent a significant burden for the head of household. This larger family (coupled with the reduced income level) might explain also the significant difference of $11,000MX in debts that Cluster 1 borrowers have over Cluster 2, which might be used to pay the additional financial commitments common in married households around its 40s, e.g. school, additional supplies, transportation, etc. (see Table 3.8 for additional readings on this cluster).

Table 3.9.B, on other hand, presents the group-specific results for Cluster 2. As we can see, the probability of default in this group is also strongly related with Age (statistically significant at more than 99%), with more senior obligors defaulting less than younger borrowers. It is important to highlight that the age defining between defaulters and non-defaulters in this cluster is slightly less than that for cluster 1 (a finding obtained from the cluster-specific CARTs below, which demonstrates how all these models are complementary in depicting a clearer picture of the defaulting behavior).

It is also interesting to note that Higher Education also presents a negative relation with PD, implying that borrower’s with a bachelor or higher degree tend to default less than individuals with basic and medium studies. This predictor is highly significant at a 95% level. Additionally, from this regression, we can see that any increase in the overall level of indebtedness (i.e. Debt Payments Monthly) and previous records as a delayed payer in Credit Bureau (captured by the binary variable Is In Credit Bureau) can predict an increase PD for Cluster 2 borrowers with almost 98% statistical confidence.
As a final remark, it is important to say that the group-specific models developed for Cluster 1 (69.19%) perform slightly better than those defined for Cluster 2 (68.52%), a relation that holds for almost all types of models (CART, RF, and Logistic regressions) and versions of the dependent variable.

As can be seen, this group-specific predictions are highly valuable because they not only allow us to discriminate more accurately between failing and non-failing applicants in each group, but also permit us to define tailored credit products and more precise commercial strategies, fostering the longer-term relationship banking that constitutes the core strategy of Mimoni Group.
CTP approach: Classification and Regression Trees (CART)

Just as in the case of Logistic regressions under CTP, our group-specific CARTS were developed based on the k-means cluster mentioned at the beginning of these section. As we saw above, Cluster 1 seems to be composed by relatively older individuals, with slightly larger families, and more level of indebtedness than borrowers in Cluster 2. This group is also characterized for having a larger proportion of married individuals than Cluster 2.

As before, our first step was to generate cluster-specific sets for the training and testing phases (the proportions in each set are shown above). Then, based on each training set, we developed a correspondent CART model for each cluster using both default and cross-validation splitting parameters. The range of c.p.'s used in our cluster-specific trees was 

\[ [0.049, 0.1] \]

for Cluster 1, and 

\[ [0.14, 0.1] \]

for Cluster 2. Finally, we evaluated the model on out-of-sample data and selected the dependent variable version with higher prediction power.

Similar to our CART findings over the whole data set, the best results in cluster-specific modeling were obtained utilizing BINFIRST_RP as our dependent variable. The combined accuracy rate of this model was 65.56%, with cluster-tailored results of 67.89% for Cluster 1 (AUC of 53%), and 65.10% for Cluster 2 (AUC of 52.10%). After optimizing for c.p., we observed an improvement in accuracies reaching 66.67% for the combined case, 68.21% for Cluster 1, and 65.77% for Cluster 2. It is noteworthy to mention that the training results provided a combined accuracy rate of 74.71%, and group-specifics of 76.61% (AUC of 67%) and 74.41% (AUC of 66.67%) for Cluster 1 and Cluster 2 respectively\(^{37}\). Figures 3.6. A and B presents the

---

\(^{37}\) As mentioned in our literature review, we could apply confidence levels in ROCs to evaluate how representative this high accuracy rates would be if the overall size of the sample and relative number of defaults is modified and overfitting is eliminated. The analysis is highly recommendable as rates of 76.61% would surpass many commercial models currently used in the industry.
group-specific CART models developed for Cluster 1 and Cluster 2 respectively (Appendix A1.7.A and B include additional trees using the same algorithm and having similar out-of-sample accuracy rates for both clusters).

Figure 3.6.A. Group-specific CART model developed for Cluster 1

From Table 3.6.A, we can see that our cluster-specific results for the CARTs developed for Cluster 1 are consistent and complementary with many of the findings in our CTP Logistic models. For instance, in our regressions, we identified how many variable patterns seem to be related with the characteristic financial dynamics of larger, married households in their early 40s. In Mexico, people in this age range tend to have sons in primary school and/or teenagers starting
junior high, which would imply additional expenses in food, books and school supplies, and transportation. Under this situation, heads of households need to be very efficient and careful utilizing their sources of financing (e.g. Maximum Credit Authorized) for otherwise this would put them rapidly in complicated financial positions. Under this logic, we would expect higher defaulting rates in borrowers with large families, low income levels and overspending and/or using almost all of their revolving credit available.

As we can see from Table 3.6.A, these variables are precisely among the best discriminators detecting borrowers with higher probability of default in Cluster 1. For example, in our tree we can observe that when a borrower with low income levels\(^{38}\) is heavily indebted (as expressed by high accumulated \textit{Monthly Debt Payments} and \textit{Debt Income Ratio} above 0.5), he will tend to present higher probability of default. Moreover, we can also see that a borrower's financial situation seems to deteriorate, and his PD seems to increase, as he approaches his credit limits, exhausting his access to alternative sources of financing (as observed in the combined path \textit{Income Declared} $\rightarrow$ \textit{Maximim Credit Limit} at the upper-right extreme of the tree).

On the other hand, from Table 3.6.B (and Appendix A.1.5.B), we can appreciate many findings for Cluster 2 that are complementary or were not fully appreciated for our group-specific regressions above. From instance, in the Logistic regressions we detected that credit bureau variables (i.e. \textit{Debt Payments Monthly} and \textit{Is in Credit Bureau}) are strong predictors of default behavior. This finding is confirmed in our tree as clearly lower level of total leverage (i.e. \textit{Debt Income ratios}), as reported in the credit bureau, are associated with lower PDs. Additionally, as

\(^{38}\) It is important to highlight that, in line with our findings about overstated income above (i.e. logs above 9), in this tree the second split (\textit{Income Declared}$\geq$8.5) still lies within the range in which people correctly states their income. Just after this differentiation is made, the leverage and financial exhausting measures come into play to discriminate between good and bad loans as we will see.

122
mentioned above, our cluster-specific CARTs allows us to grasp further insights that were not captured in our Logistic regressions. For example, from our tree here (and in A1.6.B) we can appreciate the significant role that job and residence stability have to identify clients with lower default rates. This is consistent with the demographics presented in Table 3.8 where we saw that this younger group tends to spend larger periods of time in a residence.

It is noteworthy to mention that this group has a higher proportion of individuals living in a family owned home. In Mexico, at these low levels of income, housing expenses represent a considerable proportion of an individual’s income. Therefore, it is logical to assume that when an individual displays higher residence stability (in a family owned house) this would imply less housing expenses ergo lower levels of default.

To confirm our assumption, we built a pivot table shown in Table 3.10 relative to the median length of time that a typical borrowers spends in his current residency (i.e. 14 Months for the whole 2,000 sample).

**Table 3.10. Pivot table: “Months in Current Residence vs. Cat Homeownership Family”**

<table>
<thead>
<tr>
<th>CatHomeownership Family</th>
<th>MonthsInCurrentResidence</th>
<th>Residence &lt; 14.0</th>
<th>Residence &gt;= 14.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>277</td>
<td>141</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>195</td>
<td>280</td>
<td></td>
</tr>
</tbody>
</table>

As we can see, the largest number of observations fall in the category of borrowers with long periods of time in the same residence and living in a house owned by his family (i.e. 280 obligors). This suggests that residence stability might be closely related with sharing a house with one’s relatives, possibly implying lower levels of housing expenses, and thus to more
financial flexibility to meet his loan payments. Moreover, it is interesting to note that Cluster 2 CARTs, provide an additional measure of financial discipline. As we can see from our CART model in A1.6.B, excessive uses of revolving credit used are usually associated with higher probabilities of default. The previous analysis demonstrates the finer detail we can achieve through the combined use of quantitative and data analytics methodologies. But, once again, the best results and interpretations will arise by combining these results with the qualitative expertise and industry knowledge of credit managers.

Figure 3.6.B. Group-specific CART model developed for Cluster 2
CTP approach: Cluster-specific Random Forests

As discussed before, besides the improvements in prediction power, Random Forests provide us further insights on the relative importance that certain variables have on the discriminatory process and quality of the splits. To develop these group-specific models, we first split our sample into two clusters obtain via k-means methodology, then we generated training and testing sets for each group, and finally evaluated their performance in training and out-of-sample data sets using default bucket size parameters and cross-validation optimized c.p.’s. As in our model construction over the whole set, in the CTP case, we also generated the two following diagrams per cluster: 1) number of times a variable is used for a split; and 2) GINI impurity coefficients to measure the degree of homogeneity (split quality) for each variable used.

Overall, the best results for our cluster-specific Random Forests were obtained in models using BIN_FIRST_RP as our dependent variable. The combined out-of-sample accuracy rate for this model was of 68.22% using optimal cross-validation parameters. On other hand, the cluster-specific results were of 68.01% in accuracy for Cluster 1 (AUC of 59.27%), and an outstanding out-of-sample 71.47% accuracy rate for Cluster 2 (AUC of 60.19%). It is noteworthy to mention that this last result (Cluster 2-specific RF) presented the largest out-of-sample accuracy among all CTP models and different versions of the dependent variable; moreover, it is important to highlight that the combined RF model provides the third highest result only surpassed by our “tuned” CTP Logistic Model with 68.90% accuracy, and un-clustered data improved Logistic regression with a corresponding statistic of 68.50% (see Table 3.5 for further detail on the results on other dependent variable versions).

Figure 3.7.A and B present the corresponding cluster-specific diagrams with those variables that were more often used to generate a split, this will give us an idea of the relative importance of
each variable in the classification process. Additionally, Figure 3.8.A and B display the corresponding cluster-tailored impurity diagrams with variables organized by the quality of splits they generated. As in RFs run over the whole data sample, we can see that in both diagrams and for both clusters we have groups of variables that stand out from the rest by the number and quality of splits they produce. It is noteworthy to mention that this set of variables, as we will see, is almost identical in the GINI and splits diagrams for both borrower clusters.

From Table 3.7.A and 3.8.A, we can see that the most important variables for Cluster 1 are: Age, Months in Current Employment, Months in Current Residence; Income Declared, Debt-Income ratio; Debt Payments Monthly and Maximum Credit Limit. As before, we can see that these findings are congruent with our results on cluster-specific CARTs and Logistic Models. As mentioned there, Cluster 1 seems to be characterized by relatively older borrowers (39.18 years old in average), with less financial flexibility probably as consequence of the expenses of their larger household size.

As we detected from our CTP CART models, Cluster 1 borrowers present higher leverage ratios, more accumulated debts (as reported in the credit bureau), and are approaching or are about to exhaust their Credit limit. From Table 3.7.A, we can observe that these precise predictors are the ones with the higher number of splits and GINI coefficient in Cluster 1 RF. Additionally, we must mention that an interesting insight that was not captured in our previous regressions and CARTs is the relevance of job stability, which seems to be a logical discriminator of default given the reduced financial flexibility of borrowers in Cluster 1.

On other hand, it is important to note that the list of variables per number of splits is almost identical for Cluster 1 and 2 (see Table 3.8.B), with Age again as the first and more used variable in the classification process combined with a series of job and residence stability metrics and a
group of other financial health measures. The only subtle difference for Cluster 2 is that *Months in Current Residence* substitutes *Income Declared* as the third most important variable in terms of their contribution to the purity of the partitions. These results seems to be congruent with some of the results detected in CTP CARTs for Cluster 2. As we observed there, this group has a larger proportion of individuals living in a family owned home, which could be associated with lower housing expenses, which in turn would translate in higher financial flexibility and lower probability of default (see Figure 3.6.B and A.1.6.B for examples of the relation between residence stability and PD). This could explain to some degree why Cluster 2 is the only instance where the order of the variables changes. On other hand, consistent with our findings on CTP Logistic regressions and CARTs, we can confirm that the *Debt Income ratio* and other variables of financial condition (e.g. *Debt Payment Monthly* and *Maximum Credit Limit*) also play a preponderant role to discriminate between good and bad loans for this particular group.

As a closing remark, we can observe that, as expected, the combined use of Logistic Regressions, Classification Trees, Random Forests, and clustering not only enable us to reach consistently higher predictability rates through the development of group tailored models, but most importantly allow us to demonstrate one of the core objectives of this research, i.e. that the combined use of various quantitative methods, coupled with the qualitative expertise of credit managers, can provide a significant finer picture of the dynamics behind the default behavior that would be otherwise inaccessible through the use of isolated methods or purely qualitative analysis as is the extended practice across the low-income consumer lending industry.
Figure 3.7.A. Diagram with number of times a variable is used for a split in RFs (Cluster 1)

Figure 3.7.B. Diagram with number of times a variable is used for a split in RFs (Cluster 2)
Figure 3.8.A. Impurity chart for RF model run on BIN_FIRST_RP (Cluster 1)

Figure 3.8.B. Impurity chart for RF model run on BIN_FIRST_RP (Cluster 2)
CHAPTER 4. CONCLUSIONS

Rav Hillel said: If I am not for myself, who is for me? And if I am only for myself, what am I? And if not now, when?

Babylonian Talmud

Extended interest in risk usually arise at times when the wounds of recent crisis are still fresh and then recede. But sometimes there are such major events that change completely the perspective in which we approach risks. The recent financial episodes of 2007-2008 and impressive losses in derivative markets were some of such events, putting financial institutions under exceptional scrutiny and transforming completely the way in which we manage financial credit risk.

Just a few decades ago, credit analysis consisted on qualitative assessments of an obligor’s financial health, relying mainly on the expertise and subjective understanding of the industry of credit managers. But this has changed. Nowadays, more sophisticated approaches combining the consistency, accurate prediction, and scale advantages of quantitative methods with the insightful, industry knowledge, and capacity to integrate results of experts. Today this has become the standard.

As we mentioned above, quantitative credit risk management has simply became an integral component for any company willing to survive in our highly volatile financial environment. However, although this has become the rule in “mainstream” financial areas, in low-income consumer lending the practice still rely on human-judgment, and in the best cases make just a modest use sophisticated statistical methods. As mentioned in our last chapters, this might be due to a combination of lack of structured default data bases and skepticism due to an absence of
research on the applicability and size of results that could be achieved by using sophisticated models.

One of the objectives of the present research was precisely to provide a concrete proof that the "combined" use of quantitative methods (as opposed to the prevalent practice of using mainly logistic models or other "standalone" algorithms) has direct application for the nascent low-income consumer lending industry in Latin America. Our findings demonstrate that the combined use of Logistic Regressions, Classification Trees, Random Forests, and Clustering not only allow credit modelers to reach consistently higher predictability rates over the whole sample, but most importantly that these approach are complementary enabling us to: 1) derive a significant finer detail and comprehension of the dynamics behind the default event; and 2) that the development of cluster-specific models can be a significant advantage to develop tailored products and commercial strategies. In the following lines we will discuss how our work improved upon the existing literature on Latin American low-income consumer lending in various ways.

First, this research is one of the first academic works that explored and proved the applicability of four cutting-edge quantitative methods with real operation data from one of the pioneers in the low-end credit segment in Mexico: Mimoni Group. One of the main challenges of the segment is that structured borrower information is scarce and in many cases confidential. This industry is in a nascent stage, thus information is viewed as a competitive advantage and many financial institutions are reluctant to share a large sample with enough number of default observations to generate significant results. The author acknowledges Mimoni Group's commitment to the development and diffusion of academic work, and thanks their collaboration in providing one of the largest and highest quality data bases used in the field. The results of this research would
have not being possible without their cooperation. It is important to mention that this research is one of the academic works with one of the largest numbers and proportions of defaults in these area (741 defaults and 1259 non-defaulting loans), this provided a significant advantage in terms of statistical inference as the number of failing events is what determines to a large extent the quality and power of Probability of Default Models.

Second, the close collaboration with Mimoni senior management allow us to incorporate two risk aversion outlooks (e.g. risk prone, and risk averse) based on real decision rules and criteria used by an important player in the industry. This exercise allow us to explore the consequences that two different lending attitudes can have in the prediction of PD. Moreover, given the high quality of the information we had, we could also explore the effect that partial and renegotiated due date have on the default behavior. Based on this information, we had the opportunity to define nine versions of our dependent variable (including binaries and multiclass variables) that were fundamental to explore the default phenomena under three categories: 1) models based on the original due date; 2) models considering extensions on payment deadline; 3) models incorporating the joint payment patterns under original and renegotiated due date.

After running 108 models over these nine dependent variable versions, two risk aversion outlooks, and the above mentioned payment patterns, our findings concluded that: a more risk prone treatment of the default event –i.e. default defined as delinquent loans (null payment) plus uncollectible loans (partial payment only after 90 days), with past-due loans excluded-, that considers only the payment behavior in the original date (BIN_FIRST_RP) is the best dependent variable version for the construction of models with higher accuracy rates and more consistent results across all methodologies.
Third, this is one of the first academic works for the low-income consumer lending industry making use of combined quantitative statistical methods to identify the best predictors behind higher probabilities of default in Latin America (as opposed to “standalone technique” studies). It is noteworthy to mention that our goal was not just to generate a powerful model in terms of statistical accuracy but also to obtain a better understanding of the dynamics explaining the default behavior in these segment. As we will see below, our results not only allowed us to identify a list of robust predictors of PD but also understand better the interactions among these variables.

By utilizing direct modeling over the whole sample, we reached accuracy rates as high as 71.42% for training sets, and 68.50% for out-of-sample data. Our best results under a direct modeling approach were obtained through our improved version of the Logistic regression model we generated via stepwise process. According to it, the best PD predictors are: age, number of dependents, high level of education, and marital status (singles and married).

However, as mentioned above, the complementary use of various techniques (CART and RF) and their interpretability strengths allowed us to detect two findings not accessible through the sole use of Logit. First, some borrowers seem to overstate their declared income, maybe as a means to increase their probabilities of receiving the loan. From our CART models, we detected that some obligors with declared income levels above $8,104MX consistently present higher probability of default when this income level is inconsistent with the remaining variables. Thus “inflated incomes” can be used by lenders as an effective way to detect defaulting loans. Second, through the use of Random Forest, we identified that job and residence stability play a central role in discriminating between failing and non-failing loans and that the purity of their splits presents the highest GINI levels (surpassed only by age).
On the other hand, through the development of group-tailored models (via a CTP approach) we achieved: a) consistently higher levels of accuracy across all models vis-à-vis their whole sample versions, and 2) a more granular understanding of the defaulting dynamics for the two demographic groups identified via k-means clustering. Through a detailed comparison of the means for each variable, we detected that Cluster 1 is characterized by relatively older borrowers (39.18 years old in average), with relatively larger families, more accumulated debts, and higher levels of financial leverage (less financial flexibility). On other hand, Cluster 2, in average, is comprised of younger borrowers, less indebted relative to individuals in Cluster 1, and with a larger proportion of individuals living in a family owned home, which could be associated with lower housing expenses, which in turn would translate in relatively higher financial flexibility.

In terms of accuracy our best CTP results, again, were obtained through the improved version of the stepwise Logistic regression with combined out-of-sample accuracy rates of 68.90% (highest result across all models and our two modeling approaches), and cluster-specific results of 69.19% for Cluster 1 and 68.52% for Cluster 2. It is noteworthy to mention that our training accuracy rates were as high as 77%, which would indicate that further refinement in accuracy could be achieved if an additional intermediate validation stage could be run. This would require access to larger data set and is a good next step to be included in future research.

According to our CTP improved Logistic model the best predictors of PD for Cluster 1 are: age, number of dependents, marital status, and month due. For Cluster 2, the main PD predictors are: age, overall level of past debts, delinquent record in credit bureau, and high level of education. Once again, our CART and RF results were totally consistent with this findings and also provided additional insights complementing our Logistic results.
In the case of Cluster 1, CART allowed us to identify the significant role that approaching a borrower’s credit limits have increasing the PD, given the aforementioned financial rigidity to which borrowers in Cluster 1 are subject. On the other hand, CART was fundamental to confirm the critical role that residence stability has as a predictor of lower PDs. Our CART results and additional data mining of Cluster 2 observations, allowed us to confirm that its larger proportion of individuals living in family owned homes is closely related with residence stability which us associated with lower probability of default. At these low levels of income, housing expenses represent a significant proportion of a person’s income, thus living with family could be associated with lower fixed expenses, higher financial flexibility ergo lower PDs.

Finally, it is important to mention that as in the case of direct modeling, our cluster-specific results provided complementary insights to our CART and Logistic models: 1) confirming the partitioning importance of the above mentioned variables for each cluster, and 2) identifying employment stability as an important variable in the number and quality of splits in the discriminatory process. Our objective was two-fold and we achieved it: a) we demonstrated the applicability of cutting edge-techniques in a real case, and 2) we mapped clear, statistically-supported interactions among the variables that can be immediately applied to improve our provider’s decision-making process.

In Latin America, and specifically in Mexico, the low-income consumer credit segment is experiencing an unprecedented boom (e.g. double-digits growth rates in Mexico in 2013), attracting the interest of a large number of financial players willing to capture some portion of this still under-served niche. We are convinced that the applicability of the present study is especially relevant in today’s Latin American context, and that our results clearly demonstrate that this market penetration can be accompanied by cutting-edge analytical tools to secure fair
interest rates for these underprivileged segments while still creating attractive business opportunities for lenders. After all the ideological premise of our research is to demonstrate that financial intermediaries can “do good by doing well”, and that social-driven and financial viability goals are complementary rather than contradictory.

As we write the present work, the Latin American low-income consumer lending sector is expanding as never before. We stand at an historical moment where we can develop this industry based not only on the best cutting-edge quantitative practices but on a broader business accountability that integrates the social role that financial intermediaries have in securing fair financial terms to foster the financial inclusion of the less privileged. This is not a prerogative but an obligation of true business leaders.
APPENDIX

A1.1. VIFs on Logistic regression excluding “Payments Declared”

```r
> LogModel <- glm(Dependent ~ ., data=loansDataSet1, family=binomial)
> vif(LogModel)

          GVIF Df GVIFA(1/(2*Df))
Age       1.48831 1      1.220177
Dependents 1.29190 1      1.136622
MonthsInCurrentResidence 1.23299 1      1.110404
MonthsInCurrentEmployment 1.26596 1      1.125153
IncomeDeclared 1.41945 1      1.191407
CreditHistoryLengthMonths 2.16806 1      1.472432
DebtPaymentsMonthly 1.18811 1      1.090006
RevolvingCreditAuthorized 3.29463 1      1.815114
RevolvingCreditused 2.43471 1      1.560357
MaximumCreditLimit 1.37697 1      1.173444
IsInCreditBureau 2.25874 1      1.502913
IncomeDebtRatio 1.19201 6      1.025353
SexM       1.09819 1      1.047948
CatEducationMedia 2.97424 1      1.724599
CatEducationSuperior 3.43813 1      1.854221
CatMaritalStatusCouples 3.25694 1      1.804701
CatMaritalStatusSingle 3.46530 1      1.861534
CatHomeOwnershipFamily 2.07889 1      1.441835
CatHomeOwnershipOwn 2.06482 1      1.436947
CatEmploymentIndependent 1.67543 1      1.294384
CatEmploymentLowValue 1.62436 1      1.274503
```

A1.2. A Loan Data Set used for direct approach models (Logit, CART, Random Forest).

```r
loansDataSet1 (data.frame): 2000 obs. of 26 variables:

$ Age : num 31.6 20.1 35.4 44.5 39.6 ...
$ Dependents : int 1 0 2 3 2 0 3 2 2 2 ...
$ MonthsInCurrentResidence : num 1 1 1 1 1 1 1 1 1 1 ...
$ MonthsInCurrentEmployment : num 5 1.4 0.9 0.4 13.03 ...
$ IncomeDeclared : int 5000 2500 10000 25000 13000 6000 35000 5000 900 0 ...
$ CreditHistoryLengthMonths: num 9 2 7 0 0 0 0 31 8.5 4 ...
$ DebtPaymentsMonthly: num 11.3 8.15 11.98 0 9 ...
$ RevolvingCreditAuthorized: num 0 8.54 7.21 0 0 ...
$ RevolvingCreditUsed: num 0 6.89 7.65 0 0 ...
$ MaximumCreditLimit: num 9.86 8.7 7.21 0 0 ...
$ LoanAmount: int 1000 1000 1000 1000 1000 1000 1000 1000 1000 10 ...
$ LoanTotalDue: int 1200 1200 1200 1200 1200 1200 1200 1200 1200 12 ...
$ PaymentsDeclared: num 7.6 7.35 7.6 9.83 8.92 ...
$ DebtIncomeRatio: num 0.336 0.482 0.182 0.554 0.455 ...
$ IsInCreditBureau : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 1 2 2 2 2 ...
$ MonthDue: Factor w/ 7 levels "August","December",...: 3 6 4 3 3 1 2 1 ...
$ SexM: Factor w/ 2 levels "0","1": 1 2 1 1 1 2 1 1 2 ...
```

137
A1.2.B Loan Data Set used for CPT approach controlling for multicollinearity
A1.3. Logistic regression obtained through stepwise (Dependent: BIN_FIRST_RP)

```r
> summary(step1)
Call:
glm(formula = Dependent ~ Age + IncomeDeclared + catEducationSuperior + catMaritalStatusCouples + catMaritalStatusSingle, family = binomial, data = Train)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-1.4811  -0.9189  -0.7669   1.2928   2.0589

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.059979  1.065672  -0.056    0.9551
Age          -0.046486  0.007157  -6.495  8.28e-11 ***
IncomeDeclared  0.194559  0.118918   1.636   0.101823
catEducationSuperior  -0.466237  0.129446  -3.602   0.000316 ***
catMaritalStatusCouples  -0.568572  0.204760  -2.777   0.005490 **
catMaritalStatusSingle  -0.545880  0.219490  -2.487   0.012881 *

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1784.1 on 1399 degrees of freedom
Residual deviance: 1721.3 on 1394 degrees of freedom
AIC: 1733.3

Number of Fisher Scoring iterations: 4
```

A1.4. Progressive Logistic regressions to observe variations in Income Declared coefficients

```r
> summary(LogModel1)
Call:
glm(formula = Dependent ~ Age + IncomeDeclared + catEducationSuperior + catMaritalStatusCouples + catMaritalStatusSingle, family = binomial, data = Train)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-1.5240  -0.9186  -0.7716   1.2996   2.1435

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.054347  1.089872   0.050    0.9602
Age          -0.045445  0.007155  -6.352  2.13e-10 ***
IncomeDeclared  0.189936  0.120803   1.572   0.115885
catEducationSuperior  -0.374901  0.128192  -2.925   0.003450 **
catMaritalStatusCouples  -0.568572  0.204760  -2.777   0.005490 **
catMaritalStatusSingle  -0.545880  0.219490  -2.487   0.012881 *

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1784.1 on 1399 degrees of freedom
Residual deviance: 1727.5 on 1394 degrees of freedom
AIC: 1739.5

Number of Fisher Scoring iterations: 4
```
```r
> summary(LogModel)
Call:
glm(formula = Dependent ~ Age + IncomeDeclared + catEducationSuperior, 
    family = binomial, data = Train)

Deviance Residuals:
     Min      1Q     Median      3Q     Max
-1.2993 -0.9274  -0.7929  1.3219  2.0918

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -0.992053  1.045306  -0.949  0.3426
Age                 -0.038810  0.006748  -5.752 8.83e-09 ***
IncomeDeclared      0.208947  0.119905   1.743  0.0814
catEducationSuperior -0.427711  0.125973  -3.395  0.0007 **

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1784.1 on 1399 degrees of freedom
Residual deviance: 1740.5 on 1396 degrees of freedom
AIC: 1748.5

Number of Fisher Scoring iterations: 4
```

```r
> summary(LogModel)
Call:
glm(formula = Dependent ~ Age + IncomeDeclared, family = binomial, 
    data = Train)

Deviance Residuals:
     Min      1Q     Median      3Q     Max
-1.1695 -0.9383  -0.8028  1.3428  1.9259

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.183585  0.980296   0.187  0.8510
Age                 -0.036371  0.006666  -5.457 4.85e-08 ***
IncomeDeclared      -0.08336  0.10558   -0.790  0.4300

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1784.1 on 1399 degrees of freedom
Residual deviance: 1752.1 on 1397 degrees of freedom
AIC: 1758.1

Number of Fisher Scoring iterations: 4
```

```r
> summary(LogModel)
Call:
glm(formula = Dependent ~ IncomeDeclared, family = binomial, 
    data = Train)

Deviance Residuals:
     Min      1Q     Median      3Q     Max
-0.9677 -0.9077  -0.8905  1.4628  1.5424

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.07550  0.96933   0.078  0.9380
IncomeDeclared      -0.08336  0.10558  -0.790  0.4300

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1784.1 on 1399 degrees of freedom
Residual deviance: 1783.5 on 1398 degrees of freedom
AIC: 1787.5

Number of Fisher Scoring iterations: 4
```
A1.5. Additional CARTs using default c.p. parameters.
A1.6. Algorithm used to calculate the aggregate number of times that a specific variable is
used for a split across all trees in a Random Forest.

\[
vu = \text{varUsed(ModelRF, count=TRUE)}
\]
\[
vusorted = \text{sort(vu, decreasing = FALSE, index.return = TRUE)}
\]
\[
\text{names(ModelRF$forest$xlevels[vusorted$ix])}
\]
\[
\text{dotchart(vusorted$x, names(ModelRF$forest$xlevels[vusorted$ix]))}
\]

A1.5.A. Additional cluster-specific CART model (Cluster 1)

A1.5.B. Additional cluster-specific CART model (Cluster 2)
BIBLIOGRAPHY


Instituto Brasileriro de Geografia e Estadística. 2012. Pesquisa nacional por amostra de domicílios (PNAD). Rio de Janeiro, BR.


Kurbat, M., and I. Korablev. 2002. Methodology for testing the level of the EDFTM credit measure. San Francisco: Moody’s KMV.


