



A Credit Scoring Model for the Portuguese Private Clients

Daniela Nikaitow de Oliveira



Internship report

Master in Finance



Supervised by

Carlos Francisco Ferreira Alves



2018

Banco L. J. Carregosa, S.A.

This report was made in the context of a curricular internship done at Banco L. J. Carregosa, S.A., where I have worked since September 2017 until March 2018.

Banco L. J. Carregosa, S.A. is a Portuguese credit institution, specialized in the private banking, which main goal is to advise its clients and to protect its propriety. It was founded in the XIX century, more precisely in a financial home in 1833, for the negotiation of currencies. The bank was such a novelty at that time, that was even created 13 years before the foundation of *Banco de Portugal*. 52 years later, in 1885, it was acquired by Lourenço Joaquim Carregosa, to whom the bank owns its name and a reputation of credibility and trust that remains until today. At the end of the last century, the bank originated *L. J. Carregosa – Sociedade Corretora S.A.*, that, later, was transformed into *Sociedade Financeira de Corretagem* and, finally, into *Banco L. J. Carregosa, S.A. (Banco Carregosa)*.

Nowadays, Banco Carregosa is mainly recognized for allying tradition with modernity, for the creation and development of innovative financial products and for detaining an online business. In fact, in 2000, Banco Carregosa made available the first online brokerage service in Portugal that led, in 2007, to the creation of the *GoBulling* brand.

Banco Carregosa's head office is located in Avenida da Boavista, Porto.

Biographical note

Daniela Nikaitow de Oliveira is a Portuguese girl, born in São Paulo (Brazil), in 1995, with Japanese ancestry.

In 2013, she moved from Carregal do Sal (Viseu) to Porto to enroll the BSc in Management at the School of Economics and Management of the University of Porto, which she completed in 2016.

At the same year, she enrolled in the Master in Finance program at the same institution; and, in 2017, she started a curricular internship at Banco L. J. Carregosa, S.A., which the present internship report explains the work developed and concludes the Master.

Acknowledgments

I would like to thank

- Carlos Francisco Alves, my professor and supervisor, for guiding me and helping me throughout this work, during the last several months;
- Mariana Lopes and Diamantino Leite, from Banco Carregosa's Risk Department, for receiving me in their installations and proposing me an interesting topic of research;
- my parents, Rui and Sueli, for always believing in me and investing in my education and career, to whom I am forever grateful;
- my sisters, Fernanda and Carolina, for always incentivizing me during the 23 years of my life and for all the memories shared together;
- and last, but not least, to Luís Pedro, for supporting me at all time.

Abstract

Given the increase number of bankruptcies that happened in the last years, especially after the financial crisis, and the regulatory constraints imposed by the Basel Committee on Banking Supervision and the National and European Authorities, the concern regard credit risk has increased dramatically.

This study, developed in the context of an internship done at Banco L. J. Carregosa, S.A., aims at developing a credit scoring model to calculate the probability of default of private clients, having in mind the five C's of credit: personal and socio-professional Characteristics; Character, Capital, Collateral and Cycle conditions. The data used to develop it was retrieved from a survey developed by the European Central Bank with conjunction with several countries of the Eurozone, entitled "Household Finance and Consumption Survey", in 2013.

The research evidences that what seems to play a major role when evaluating credit scoring models is the value of the cut-off; and that it is better to regress a model individually for each country (instead of combining information of different countries and benefiting from a higher number of observations). The model proposed presents a total accuracy rate equal to 78.29% and better accuracy results than the probabilistic model developed by Henriques (2014) and the rating model developed by Saunders and Cornett (2012).

Key-words: Credit scoring model, credit risk, probabilistic model

JEL-Codes: C51, D14, E51, G21

Resumo

Tendo em consideração o aumento do número de falências que ocorreram nos últimos anos, especialmente após a crise financeira, e as mudanças em termos de regulação impostas pelo Comité da Basileia na Supervisão Financeira, pelos supervisores nacionais e Europeus, a preocupação em relação ao risco de crédito tem aumentado drasticamente.

Desta forma, este estudo, desenvolvido em contexto de estágio curricular realizado no Banco L.J. Carregosa, S.A., tem como objetivo o desenvolvimento de um modelo de crédito para aferir a probabilidade de incumprimento de clientes particulares, tendo em conta os cinco C's do crédito: Características pessoais e socioprofissionais, Carácter, Capital, Colateral e Condições da economia. A informação utilizada para desenvolver o modelo foi retirada de um inquérito desenvolvido pelo Banco Central Europeu juntamente com diversos países da zona euro, em 2013, intitulado “Inquérito à Situação Financeira das Famílias” (ISFF).

Este estudo proporciona evidência de que o que causa maior impacto quando este é avaliado, é o cut-off escolhido. Para além disso, é importante fazer a regressão de um modelo usando informações individuais do país em causa, em vez de se usar informações de vários países, apenas para fazer proveito de um maior número de observações. O melhor modelo apresentado neste estudo apresenta uma taxa de acerto global igual a 78.29%, que são resultados melhores que os alcançados por Henriques (2014) e Saunders and Cornett (2012), no seu modelo de rating.

O modelo desenvolvido pode ser utilizado por qualquer instituição financeira, que beneficiará de um modelo único, desenvolvido com informação providenciada pelo Banco Central Europeu e pelo Instituto Nacional de Estatística.

List of Contents

Chapter 1: Introduction	1
Chapter 2: Literature Review	5
Part A.	5
1.1 Corporations vs. Retail Loans.....	5
1.2 Traditional Approaches to Credit Risk.....	6
1.2.1 Expert Systems	6
1.2.2 Rating Systems	7
1.2.3 Credit Scoring Models	9
1.3 BIS Basel New Capital Accord.....	13
Part B.....	14
Chapter 3: Data Description & Methodology	19
Part A: The Survey	19
Part B: Methodology	20
Part C: Data Description.....	28
Chapter 4: The Model.....	39
Part A: The Model	39
Part B: Comparison with other models	45
1. Henriques (2014)' Model – Version 1 and 2	45
2. Model by Saunders and Cornett (2012)	47
Chapter 5: Application of the model on other European countries	53
Chapter 6: Conclusions	57

References.....	59
Annexes.....	63

List of Tables

Table 1: Different methods to construct a credit scoring model and respective technique and summary Source: Anderson (2007)	11
Table 2: Overall accuracy of the models developed by the authors.....	16
Table 3: Main variables included in the models of the mentioned authors	18
Table 4: Variables that have survived Test 1	24
Table 5: Variables that have survived Test 3.1 and Test 3.2 *: Test 3.2 was computed given the acceptance of the null hypothesis on Test 2. **: Test 3.1 was computed given the rejection of the null hypothesis on Test 2.	25
Table 6: Binomial variables that have survived Test 4.....	26
Table 7: Model 0 *: p-value < 0.1 **: p-value < 0.05 ***: p-value < 0.01	41
Table 8: Model A and model B *: p-value < 0.1 **: p-value < 0.05 ***: p-value < 0.01	43
Table 9: Accuracy of the model A with a 15% cut-off	44
Table 10: Accuracy of the model B with a 15% cut-off	44
Table 11: Model developed by Catarina Henriques (2014)' model – version 1 and 2 *: p-value < 0.1 **: p-value < 0.05 ***: p-value < 0.01 n/a: information not available.	46
Table 12: Accuracy rates of Model A, Model B, Henriques (2014)'s Model – Version 1 and Henriques (2014)'s Model – Version 2, with a cut-off equal to 15%	47
Table 13: Variables, values and weights of the rating model developed by Saunders and Cornett (2012)	49
Table 14: Accuracy of Saunders and Cornett (2012)'s model with the conversion of the variable “total gross income” from EUR to USD, with a range between 120 and 190.....	49
Table 15: Accuracy of Saunders and Cornett (2012)'s model with the adjustment of the variable “total gross income” using PPP with a range between 120 and 190.....	50
Table 16: Frequency of the scores from the model of Saunders and Cornett (2012) after the adjustment of the variable "total gross income"	51
Table 17: Model C *: p-value < 0.1 **: p-value < 0.05 ***: p-value < 0.01	55

Table 18: Accuracy of model C with a 15% accuracy rate, without discriminating the data of the countries	55
--	----

List of Annexes

Annex 1.: Market share calculation based on 2015 applicable turnover from credit rating activities and ancillary services in the EU (European Securities and Markets Authority, 2016).	65
Annex 2.: Initial 68 variables considered.....	66
Annex 3.: Variables which outliers were controlled, and respective minimum and maximums (before and after the winsorization process) and respective percentage of winsorization. ..	68
Annex 4.: Variables tested and respective results (the ones in red are the ones that were automatically excluded due to their results in any one of the tests or for not being available).	70
Annex 5.: Distribution of the variable “situation at current job”.....	76
Annex 6.: Distribution of the variable “Sector of the company where it has main job”.....	77
Annex 7.: Distribution of the variable “Year of the acquisition of the main residence”.....	78
Annex 8.: Accuracy of the models A and B, using cut-offs equal to 50%, 30%, 20% and 10%.	79
Annex 9.: Accuracy of Catarina Henriques (2014)’s model, using cut-offs equal to 50%, 30%, 20%, 15% and 10%.	82
Annex 10.: Accuracy of Catarina Henriques (2014)’s regressed model, using cut-offs equal to 50%, 30%, 20%, 15% and 10%.	84
Annex 11.: Accuracy rates of Saunders and Cornett (2012)’s model, using ranges between 240 and 310; 250 and 320; 260 and 330; 270 and 340; and 280 and 350.	86
Annex 12.: Accuracy rates of model C with aggregated data from Portugal, France, Italy and Spain, for cut-offs equal to 50%, 30%, 20% and 10%.	88

Annex 13.: Accuracy rates of model C discriminating the data from each country (Portugal, France, Italy and Spain), for cut-offs equal to 50%, 30%, 20% and 10%.....	90
Annex 14.: Output of model C when regressing individually for each country; and respective accuracy rates, for Portugal and Spain, for cut-offs equal to 50%, 30%, 20%, 15% and 10%.	97

List of Abbreviations and Acronyms

ANN	Artificial Neural Networks
BCE	Banco Central Europeu
BIS	Bank for International Settlements
DA	Discriminant Analysis
DF	Degrees of Freedom
DT	Decision Trees
ECB	European Central Bank
HFCS	Household Finance and Consumption Survey
INE	Instituto Nacional de Estatística
IRB	Internal Ratings-based Approach
ISFF	Inquérito à Situação Financeira das Famílias
Logit	Logistic Regression
LR	Linear Regression
NAIC	National Association of Insurance Commissions
OECD	Organization for Economic Co-operation and Development
PD	Probability of Default
Probit	Probabilistic Regression

Chapter 1:

Introduction

Financial institutions, in their daily activities, perform the indispensable function of channeling funds from people that have surplus funds (suppliers of funds) to those with shortage of funds (users of funds), through credit. This process starts with the initial loan application and ends with the successful repayment of the loan or its default. Due to asymmetric information, the default is hard to predict because who borrows money has always more information than the one who lends (Kocenda & Vojtek, 2009). Uncertainty also makes complicated to forecast who will default and who will repay the loan. Although retail lending is one of the most profitable investments in a lender's asset portfolio. The increase number of conceded loans also increases the number of defaulted ones. This arises a risk commonly known as *credit risk*. It exists since the existence of lending itself, back as far as 1800 B.C.E¹. and the concept has been the same since the ancient Egyptian times (Caouette, Altman, & Narayanan, 1998). Credit risk is the risk that a borrower may not repay a loan, because it is not able or unwilling to, which means that the lender may lose the principal and/or the interest associated with it. This risk arises because it is not possible to ensure that the borrowers will pay back the amount borrowed. According to Obrova (2012), credit risk can also be called "*loan risk*" and Caouette et al. (1998, p. XV) defines credit as being "*nothing but the expectation of a sum of money within some limited time*" and, consequently, defines credit risk as "*the chance that this expectation will not be met*". There is credit risk anytime someone takes a service or a product, without paying immediately for it.

Over the last decades, credit risk measurement had to evolve radically, due to a number of reasons. According to Altman and Saunders (1997); Caouette et al. (1998); Hand and Henley (1997) some of the reasons include: (i) a worldwide increase in the number of

¹ B.C.E. means "*Before Common Era*", known by many as B.C., "*Before Christ*".

bankruptcies, translating into a more concern regarding credit risk; (ii) a trend towards disintermediation by the highest quality and largest borrowers, by investing directly in the money markets; (iii) the increase of competition; (iv) a declining value on real assets, translating in a decrease value of collaterals; (v) the drive for diversification and liquidity; (vi) an increase growth of off-balance sheet instruments with inherent default risk exposure; and (vii) regulatory changes, such as the requirements created by the Basel Committee on Banking Supervision². Happily, in the last two decades, it had become easier to develop risk measurement approaches due to the development of technology and the availability of information through the World Wide Web. Banks need to make use of this increasing sophistication in terms of techniques, strategies and scientific and mathematical models to measure the credit risk of loans in order to price them correctly; and to set appropriate limits on the amount of credit extended to a client. As Caouette et al. (1998) state, managing risk is the same as creating a custom-made suit: it is crucial to measure the customer's needs and capacities to make sure the financing is a good fit. This is very important because the default of a singular borrower can have a significant impact on the value and reputation of the financial institution. According to Constangioara (2011, p. 162), there is an urgent need to develop methodologies to assess credit risk since the development of the markets has led to "*over-indebtedness and consumer bankruptcy phenomena*", especially after the financial crisis of the last decade. Thanks to this, academics and practitioners have started developing new and more sophisticated credit scoring systems and models to protect both the lenders and the good borrowers (which potentially will access to better conditions as lower is the rate of default of the other clients).

Having this in mind, and considering the fact that I was an intern at Banco L. J. Carregosa, the idea of this study is to develop a credit-scoring model to assess the creditworthiness of private clients of the Portuguese banking industry, considering their default probability, having in mind a work previously done by Henriques (2014) and to overcome her results. It is important to develop such model, because, in the United States (U.S.), it is used the FICO Score in 90% of lending decisions (Sousa, Gama, & Brandão,

² The Basel Committee on Banking Supervision is a committee of banking supervisory authorities, which goal is to provide a forum for regular cooperation on banking supervisory matters, to enhance understanding of key supervisory issues and to improve the quality of banking supervision worldwide. It was established in 1974.

2016) but, in the OECD³ countries (where Portugal and many European countries are included), banks follow the approach proposed by the Basel Committee in which each bank is encouraged to develop its own internal scoring model (Bank for International Settlements, 2006). In order to do so, it is important to collect a data set, which will be the responses to a survey made by the European Central Bank (ECB) - the European Household Finance and Consumption Survey (HFCS)– in conjunction with several countries on the European Union, including Portugal. This survey provides sociodemographic and finance information about households that is indispensable to the creation of a good retail credit-scoring model.

The lack of retail models in the industry is mainly due to the scarce information about households (because they are informationally opaque and borrow relatively infrequently (Kocenda & Vojtek, 2009)); the costs associated with retrieving such information; and the difficulties that banks face when trying to access the existent databases. Hence, with the development of the credit model for retail banking, we think this study will be useful to the banking industry of the European countries since it may be used by any financial institution that feels it is appropriate to its business.

The rest of the study proceeds as follows: chapter 2 presents the literature review; chapter 3 provides a comprehensive description of the data and the methodology followed; chapter 4 presents the model developed, its analysis, and a comparison with other models in the literature; chapter 5 presents an implementation of the model developed on other countries, namely France, Italy and Spain; and, at last, chapter 6 presents the conclusions and suggestions for future research.

³ OECD translates to Organization for Economic Co-operation and Development, which is an intergovernmental economic organization with 35 country members, founded in 1960, in order to stimulate economic progress.

Chapter 2:

Literature Review

Part A.

1.1 Corporations vs. Retail Loans

The focus of this study, as previously mentioned, is to identify, develop and compute a credit model for private clients of the banking industry. Since *“financial institutions manage credit risks for business and consumers differently”* (Šušteršič, Mramor, & Zupan, 2009, p. 4736), it is relevant to make a small distinction between lending to corporations and lending to individual borrowers. The Bank for International Settlements (2001, p. 55) (BIS)⁴ defines retail credit as *“homogeneous portfolios comprising a large number of small, low value loans with either a consumer or business focus and where the incremental risk of any single exposure is small”*. These types of loans include loans made to individuals, such as credit cards, residential mortgages and home equity, auto or educational loans (Allen, DeLong, & Saunders, 2004). The differences between corporate and retail loans rely on the amount lent to each one of them, being much smaller to retail; and, while for corporate loans various financial ratios are used to construct models to assess credit risk or the probability of default (PD), like the z -score developed by Altman; in retail banking, various sociodemographic characteristics are collected to make a proper decision about the client. Moreover, since lenders face fixed costs when lending, lending to individuals becomes more expensive per dollar lent. Another disadvantage of lending to small firms or individuals is the lack of information since they tend to be more informationally opaque. Their information is not public.

⁴ The BIS is an international financial organization owned by 60 member central banks, headquartered in Basel, Switzerland. It was established on 17 May 1930, and its mission is to serve central banks in their pursuit of monetary and financial stability.

Despite these disadvantages, it is still important to pay attention to credit conceded to individuals. According to statistics of Banco de Portugal, discounting the numbers to December 2016, the credit stock conceded to individuals in Portugal amounted €125 billion, of a total of €203 billion. As it can be seen, €125 billion is a huge number and, in percentage, refers to 61.58% of the total credit conceded by the financial sector (Banco de Portugal, 2017). Moreover, in the first nine months of 2017, the amount of credit conceded to consumer credit amounted to €17.7 million per day, a 12% increase in homolog terms. This increases the concern that Banco de Portugal has in relation to credit risk since it fears that households are falling into a “*spiral of indebtedness*”, again (Soares, 2017).

1.2 Traditional Approaches to Credit Risk

As Allen et al. (2004); Altman and Saunders (1997); Hand and Henley (1997) among others, state, in the last 30 years, some methodologies to assess credit risk among financial institutions were developed. The traditional ones focus on estimating the PD's, including the probability of a bankruptcy filing, default or liquidation. According to the BIS, a client is in default if it is more than 90 days overdue with a payment connected with the loan; and, according to Banco L. J. Carregosa (2017), the default takes place when a payment is not made at the predetermined date.

Some examples of these traditional models include expert systems (where artificial neural networks can be included); rating systems; and credit scoring models.

1.2.1 Expert Systems

Expert systems rely on the subjective capacity of professionals in assessing the likelihood of default, according to some personal characteristics. Individuals become experts over the course of their careers, gaining authority as they acquire experience and demonstrate skills (Caouette et al., 1998).

One prominent example of such systems is the 5 C's of credit: character, capital, capacity, collateral, and cycle. The first one, character, is related to the reputation of the potential borrower. It is a measure of the borrowers' willingness to repay and his/her repay history. The second one, capital, is the leverage of the borrower. Capacity concerns the ability to repay, which reflects the volatility of the borrower's earnings. Regarding collateral, it means that a banker has claims collaterals pledged by the borrower. The collateral depends on the PD that the professional believes the borrower has. Finally, the cycle conditions refer to the state of the business cycle. This last “C” is very important because a client, that seems

to be very independent of the state of the economy, may be affected by economic downturns and financial crisis (Allen et al., 2004; Altman & Saunders, 1997; Gonçalves, Gouvêa, & Mantovani, 2013).

In order to develop a more objective expert system, the artificial neural networks (ANN) have been introduced. Basically, an ANN uses historical repayment experience and default data to assess the PD of a client. Each time the network evaluates the credit risk of a new loan opportunity, it updates the data in order to “*continually learns from experience*” (Allen et al., 2004, p. 734). This feature makes the ANN a system very flexible and adaptable (Abdou & Pointon, 2011; Altman & Saunders, 1997) and it works due to the development of technology and the appearance of new methodologies, like artificial intelligence.

Since the network fits a system of weights to each financial variable included in the database, the downturn of the methodology lies on the fact that “*too much training*” may result in poor out-of-sample estimates. This can happen because the network may be “*over fit*” to a particular database (Allen et al., 2004), losing its universal characteristic. Allen et al. (2004) also underline the fact that it is very costly to implement and maintain this methodology, it is a slow procedure, and it may miss transparency through the process.

1.2.2 Rating Systems

A rating system was born to answer the question “*How do lenders determine the creditworthiness of potential borrowers and assure themselves of the continued soundness of borrowers after a loan has been extended?*” (White, 2002, p. 44). In order to answer the question, financial intermediaries may develop the necessary information themselves to construct a rating system or may turn to credit rating specialists, known as *Credit Rating Agencies*. These agencies can help those who cannot create rating systems themselves, by eliminating asymmetric information that surrounds the lending relationships.

A firm’s credit rating is a measure of the firm’s propensity to default. Credit ratings provide individual and institutional investors with information that assists them in determining whether issuers of debt obligations and fixed-income securities will be able to meet their obligations with respect to those securities.

Internal credit ratings are a progressively more important element of credit risk management. Within the past few years, the credit-related businesses have become gradually more complex and the number of counterparties has grown rapidly. Thanks to this, many banks, especially the bigger ones, have introduced more structured and formal systems for approving loans, portfolio monitoring, and management reporting. Internal ratings are

crucial inputs to all such systems as well as to quantitative portfolio credit risk models, like the one proposed by the Basel Committee.

Just like a public credit rating produced by credit rating agencies such as Fitch Ratings, Moody's or Standard & Poor's, a bank's internal rating summarizes the risk of loss due to failure by a given borrower (Treacy & Carey, 2000). The main difference between the ratings produced by agencies and banks rely on the fact that internal ratings are assigned by bank personnel and are usually not revealed to outsiders, due to competitive advantage issues.

The National Association of Insurance Commissioners (NAIC)⁵ requires companies to rank their assets according to six different classifications corresponding to the following credit ratings: A and above, BBB, BB, B, below B and default. But, currently, the specifics of internal systems vary across banks. Each one assigns grades and its associated risk according to their needs and typical clients (Allen et al., 2004).

The drawback of this credit assessment methodology relies, mainly, on its complexity. In order to develop an internal rating system, considerations about costs, efficiency of information gathering, consistency of ratings produces, and staff incentives must be made (Treacy & Carey, 2000).

1.2.2.1 Rating Agencies

Credit rating agencies (such as Moody's Investors Service; Standard & Poor's Corporation; or Fitch Ratings) provide investors a forward-looking opinion on the relative credit risks of financial obligations, such as interest, preferred dividends, repayment of principal, insurance claims or counterparty obligations (Fitch Ratings, 2017; Moody's Investors Service, 2017). It is their job to inform investors about the likelihood of them receiving their money back, as scheduled for a given security. Despite what many may think, it is not their job to make recommendations about buying or selling; their job is only to express informed decisions about creditworthiness, through independent, objective, transparent and high-quality analytic processes (Caouette et al., 1998). This does not mean, however, that, in the theoretical approach, credit ratings should be exclusively attributed by a commercial rating agency. In fact, many major financial institutions maintain their own credit rating systems, based on internally developed methodologies (internal ratings), as it was already mentioned. Moreover, just because these agencies are specialized in attributing

⁵ The NAIC is the U.S. standard-setting and regulatory support organization. It establishes standards and best practices, conducts peer review and coordinates the country regulatory oversight.

ratings, that does not mean that they are accurate. The rating is just an opinion. As Fitch Ratings (2017, p. 4) states, “*ratings are not facts and, therefore, cannot be described as being «accurate» or «inaccurate»*” and “*users should refer to the definition of each individual rating for guidance on the dimensions of risk covered by such rating*”.

Despite that, rating agencies are especially important for borrowers, since they facilitate their access to new markets and diminish the costs of their borrowings. Individuals with no expertise in financial markets can easily enter the market by buying the services from these agencies.

Nowadays, the three biggest players are Fitch Ratings, Moody’s Investors Services and Standard & Poor’s (S&P). These three rating agencies provide extensive rating coverage in Europe, especially Moody’s and S&P. Despite the existence of more than 30 other rating agencies in Europe, these three dominate the market with a market share of more than 90% (see Annex 1).

Each one of these agencies uses a system of alphanumeric letter grades to allocate the issue or issuer on a spectrum of credit quality. The spectrum goes from AAA/Aaa (very low probability of defaulting or a strong capacity to meet financial commitments) to C/D (very high probability of defaulting). The higher the grade, the higher is the probability that principal and interest payments will be paid. The debt rated Baa3/BBB- or above is considered to be of investment grade quality; while issues rated below Baa3/BBB- are viewed as speculative and risky.

Recently, on September 2017, Portuguese Republic’s credit rating was restored to investment grade by S&P, going from BB+ to BBB-; and by Fitch Ratings, going from BB+ to BBB+, on December 2017. It was BB+ since 2012 when the country was going through a bailout program provided by the European Union and the International Monetary Fund (Lima, 2017). This means a lot to Portugal. As the current Portuguese Finance Minister, Mario Centeno, states:

[The upgrade of the country’s credit rating] “*(...) allows a much vaster array of investors to have Portuguese debt in their portfolios. It also allows private debt to benefit from these better financing conditions, and this is very relevant for Portuguese banks*” (Lima, 2017).

1.2.3 Credit Scoring Models

A credit scoring model is “*the term used to describe formal statistical methods used for classifying applicants for credit into good and bad risk classes*”, as states Hand and Henley (1997, p. 523) and it is considered as “*one of the most successful applications of statistics and operations research*” (Crook,

Edelman, & Thomas, 2007, p. 1448). According to Thomas (2000, p. 151), “*credit scoring is essentially a way of recognizing the different groups in a population when one cannot see the characteristic that separates the group but only related ones*”. According to the same author, this idea was first introduced by Fisher, in 1936, and then developed by Durand, in 1941, who was able to recognize that the separation of classes was useful to separate among good and bad loans.

Although credit risk is more than 5,000 years old, credit scoring models have just a little more than 50 years (Samreen and Zaidi, 2012). The first one appeared in the 1950’s when the first consultancy of credit risk was formed by Bill Fair and Earl Isaac (Baker & Filbeck, 2013). In the late 1960’s, with the development of credit cards and with the need for more automatic decision-making processes, banks and some credit cards issuers realized the importance of credit scoring models (Thomas, 2000). Only some years after, the use of credit scoring techniques was extended to other products, like home loans and small business loans (Thomas, 2000). In the 1980s, with the development of technology science, new methodologies were developed to compute more advanced scorecards, like logistic regression and linear programming. More recently, artificial intelligence techniques, like neural networks, appeared (Thomas, 2000). The first banks to use scoring models for small businesses were mainly big banks that had at their service historical loan data to build a robust model, like *Hibernia Corporation*, *Wells Fargo*, *BankAmerica*, *Citicorp*, *NationsBank*, *Fleet* and *Bank One* (Mester, 1997).

Statistical models, also called *score-cards*, were developed through the years and they “*use predictor variables from application forms and other sources to yield estimates of the probabilities of defaulting*” (Hand & Henley, 1997, p. 524). The decision to whether grant or not credit is made comparing the PD with a predefined threshold. Nowadays, standard statistical models include discriminant analysis (DA), linear regression (LR), logistic regression (logit), probabilistic regression (probit) and decision trees (DT) (Constangioara, 2011; Costa & Farinha, 2012; Hand & Henley, 1997). The two most used ones are the logit and the DA, which was pioneered by Altman in 1968 (Allen et al., 2004). The downturn of the DA relies on the fact that assumes linearity between variables, which is not always true. On the other hand, the logit is better because do not require the multivariate normality assumption (Šušteršič et al., 2009).

Table 1 summarizes the methods previously mentioned:

<i>Method</i>	<i>Main technique</i>	<i>Summary</i>
<i>Linear regression</i>	Ordinary Least Squares	Determine formula to estimate continuous response variable.
<i>Discriminant Analysis</i>	Mahalanobis distance	Classify cases into prespecified groups, by minimizing in-group differences.
<i>Logistic Regression or Probabilistic Regression</i>	Maximum likelihood estimation (MLE)	Determine formula to estimate binary response variable.
<i>Decision trees</i>	RPA's	Uses tree structure to maximize group differences. Complex for large trees.
<i>ANNs</i>	Multilayer perception	AI technique, whose results are difficult to interpret and explain.
<i>Linear programming</i>	Simplex method	Operation research technique usually used for resource allocation optimization.

Table 1: Different methods to construct a credit scoring model and respective technique and summary
Source: Anderson (2007)

All these different models use financial variables that are believed to have statistical explanatory power in differentiating defaulting firms from non-defaulting and sociodemographic variables to assess the possibility of having individual clients defaulting. The variables can be related to the client's stability, like time at current address and/or job; regard financial sophistication, like the possession of checking accounts, savings accounts, credit cards and time at the current bank; or related to the consumer's resources, like his/her ownership status, employment and number of children (Obrova, 2012, p. 661). However, characteristics such as race, religion, national origin, gender, color or marital status cannot be used in the U.S.⁶ and should not be used due to racism and prejudice. After the parameters of the model are assessed, the loan applicants are assigned a score that classifies the loan as good or bad, that can be, consequently, converted into a PD.

According to Mester (1997), 97% of banks use credit scoring for approve credit cards applicants; and 70% use it to their small business lending.

The credit scoring has the advantage that a loan can be conceded independently of its location since the process can be done without a face-to-face contact. Documentation is minimal; it is inexpensive to implement, without subjection like the expert models (Allen et al., 2004). But, on the other hand, data limitations, the so called "*population drift*", sample bias

⁶ This is stated at the Equal Credit Opportunity Act (ECOA), created in the U.S. in 1974.

and the assumption of linearity are the downturns of this methodology (Allen et al., 2004; Altman & Saunders, 1997; Hand & Henley, 1997).

Despite what happens among European countries, in the U.K. and in the U.S. people are being credit scored or, as Thomas (2000) states, “*behavior scored*”, at least once a week. This is mainly done through the “*FICO model*” and it aims to monitor the clients’ propensity to default.

1.2.3.1 FICO Model

The most used credit scoring model today is the one developed by Fair, Isaac and Co. Inc. – the *FICO model*. This model was specially developed to meet the needs of individual costumers, who needed credit. Over the years, the model was developed to cover other business areas, such as to evaluate credit of small businesses, including trade credit (*CrediFYI.com*) or loan credit (*LoanWise.com*). In 2001, the original *FICO model* was improved and costumers could determine their credit score using the internet, through the website *myfico.com*.

As there is the FICO score, there are other credit scores across banks and firms. Usually, the differences between them are the variables that compose the model. For example, the FICO score uses variables related to credit history and credit reports to determine a score that goes from 300 to 850. The authors of this score choose to not include variables that are capable of bias a lender, such as race, religion, national origin and marital status (Allen et al., 2004).

The FICO score and scores alike exist mainly in the U.S.A. and in the U.K.. It is not a methodology usually followed by European banks. This happens due to three different reasons. First, there is lack of information about households, since they are informationally opaque and do not have their own information public, which complicates the creation of a robust model. Second, despite the existence of some surveys made to households about their financial stability, banks face many difficulties when trying to access them. At last, even if banks had all the information that was needed to create such models, there are costs associated with the creation of a credit scoring model. Since individuals borrow less money when comparing to big clients and corporations, it becomes more expensive, per dollar/euro lent, to create a good credit scoring model to individual clients (Kocenda & Vojtek, 2009).

1.3 BIS Basel New Capital Accord

The Basel Committee on Banking Supervision is an important player when concerning the financial risk regulation network, by setting risk management regulations to financial institutions all over the world. It was established in 1975 by the Central Bank Governors of the Group of Ten (G10) countries, with representatives of 13 different countries (Belgium, Canada, France, Germany, Italy, Japan, Luxemburg, the Netherlands, Spain, Sweden, Switzerland, the United Kingdom and the United States); and meets regularly in Basel, at the Bank for International Settlements.

The (first) Capital Accord (*Basel I*) was released in July 1988, in order to establish a minimum capital standard to protect financial institutions against credit risk. In 1993, the market risk was included in the scope of the accord. In 1998, the accord was fully reviewed in order to take into account all risks faced by financial institutions, including the operational risk, and, thanks to that, a new Basel accord was created – the *Basel II*.

The proposed Basel New Capital Accord allows banks to choose which approach they prefer when determining their capital requirements – capital that is set aside to cover unexpected losses. Regarding credit risk, there are two approaches that banks can follow: the Standardized Approach, which is a standardized manner to assess credit risk, supported by external credit assessments (like Rating Agencies); and the Internal Ratings-based Approach (IRB), that allows banks to use their own internal rating system (subject to prior approval by the National Supervisor) (Bank for International Settlements, 2006).

White (2002) criticizes the proposal by the BIS, by saying that it only creates demand to rating agencies and do not designate how credit rating firms should be certified. This happens because, in order to the Standardized Approach to be effective, banks can only rely on credit ratings by firms that are certified – ECAPs (External Credit Assessment Institutions). Moreover, as White (2002, p. 56) states, “*adoption of the BIS proposal in its current form is thus likely to raise worldwide barriers to entry into the credit rating industry*”, since it is only advantageous, for the rating firms, if they can be certified, otherwise they would lose a relative amount of possible clients. In relation to the IRB approach, Crook et al. (2007) believe that big banks tend to choose this approach because it allows banks to have less capital, earning higher returns on equity, since they are more or less free to choose the model to be used.

Part B.

In order to develop a credit-scoring model or any model, some steps must be chronologically followed. First, it is important to collect information about the population. Some surveys are available to research, like the one that will be the base of this study - the *Household Finance and Consumption Survey* (2013), inquired by the European Central Bank (ECB). Secondly, it is fundamental to investigate which type of model will produce the best results to the objective in question: LR, DA, logit, probit, ANNs, among others; and which set of variables to include in the model. Then, the model must be run and some tests to infer its significance and adequacy to the purpose in question must be made. Only after going through these steps, it is possible to assess if the model developed was made properly and is adequate to the final objective that it is to assess the creditworthiness of retail clients of the European banking industry.

West (2000) believes that ANNs perform better when assessing the creditworthiness of clients, but that the logit is a good alternative. In order to evidence that, the author conducted a study using two databases – German and Australian credit data – to assess which models and types of models are more suitable: parametric models (like DA and logit), nonparametric methods (like k nearest neighbor and kernel density), DT's or ANNs. The author concluded that ANNs models can increase the credit scoring accuracy from 0.5 to 3%, which can save millions to the financial institution; that the best ANNs to assess the creditworthiness of clients are *mixture-of-experts* and *radial basis function neural networks*; and that the logit is indeed a good alternative, since the difference in terms of accuracy is very small when comparing to ANNs.

Šušteršič et al. (2009) created a credit scoring model using ANNs. Using a data set provided by a Slovenian bank with internal bank data available for 581 short term consumer loans in the period of 1994 to 1998, and comparing with a logit model, the authors came to the conclusion that EBP ANNs (the type of ANNs used) have the best accuracy and the lowest value for error type II, with 79.3% of accuracy, 17.8% error type II and 29.9% error type I. The main objective of this study was to conclude about the variable selection method used, which was a principal component analysis and a genetic algorithm (*Kobonen neural network* and random method). The model started initially with 67 variables and ended with only 21. The author chose to make a comparison between ANNs and LR because “*the logit model is the most promising and widely used statistical credit scoring model*” (Šušteršič et al., 2009, p. 4750).

On the same line of research, Imtiaz and Brimicombe (2017) conducted a study to verify which model is the best to assess the creditworthiness of clients when imputation technique⁷ is used and when it is not. The authors concluded that ANNs present better results when the imputation technique is applied, since it increases the availability of data and, therefore, increases the accuracy rate of classification of ANNs. In the absence of the technique, the author concluded that, despite having DT's performing better when training the model, ANNs performed better when the model was tested. Despite the overall better accuracy of ANNs models, its downturns rely on the fact that it takes too much time to train the model when there is a big sample, which is when the model presents its better results. Moreover, according to the author, and in the context of risk control, it is more meaningful to test the client risks without imputation, since it can bias the sample.

Samreen and Zaidi (2012) conducted a study to assess which type of model produced the better results when assessing the creditworthiness of Pakistan's clients. The author interviewed 250 clients of the banking industry of Pakistan and concluded that the logit regression had an accuracy rate of 98.8% and the DA for individuals presented an accuracy rate equal to 95.2%. The variables used by Samreen and Zaidi (2012) included sociodemographic variables, such as marital status, age, number of dependents, occupation, working period with the last and current employer, and monthly net income; and finance related variables, such as loan tenure, loan period, banking references at the bank, credit history and loans from others banks.

Table 2 summarizes the different conclusions, in terms of total accuracy, that the different authors determined, as well as the technique used to assess the accuracy – *AUC* or *Error Rate*.

<i>Author (year)</i>	<i>Logit</i>	<i>ANNs</i>	<i>DA</i>	<i>CT's</i>	<i>Database</i>	<i>Obs.:</i>
<i>(West, 2000)</i>	76.30%	77.57%	72.60%	69.56%	From German	<i>Accuracy technique used: Error Rate</i>
	87.25%	87.61%	85.96%	84.38%	From Australia	<i>Accuracy technique used: Error Rate</i>
<i>(Šušteršič et al., 2009)</i>	76.10%	79.3%	-	-	From Slovenia	Selection variable technique
	71.30%	73.00%				

⁷ Imputation technique is a technique that it is used when there are missing values in the sample, by replacing missing values with substitute data. It presents some advantages like avoids the decrease in the number of values of the sample that it is studied; but may introduce bias and reduce efficiency.

	72.00%	70.70%				Accuracy technique used: Error Rate
(Samreen & Zaidi, 2012)	98.80%	-	95.20%	-	From Pakistan	Accuracy technique used: Error Rate
(Constancioara, 2011)	96.00%	96.00%	-	96.00%	From Hungary	Stepwise selection Accuracy technique used: Error Rate and AUC
(Kocenda & Vojtek, 2009)	86.40%	-	-	83.00%	From The Czech Republic	With and without "Own resources" Accuracy technique used: AUC
(Imtiaz and Brimicombe, 2017)	90.29%	90.99%	-	89.57%	From Taiwan	Without imputation technique Accuracy technique used: Error Rate and AUC

Table 2: Overall accuracy of the models developed by the authors

As it can be seen from the previous table, ANNs models seem to be the most accurate ones but just with a minimal difference from logit regressions. Despite the fact that DA is one of the most used ones, its accuracy is not that great when comparing to other models. The main reason why DA is still one of the most used models today relies on the fact that institutions developed DA models in the past and are now reluctant to develop better models, due to the costs associated with it and the time it consumes. According to Hand and Henley (1997, p. 535), there is no best model. It depends on “the data structure, the characteristics used, the extent to which it is possible to separate the classes using those characteristics and the objective of the classification”. And a model to be considered as “best”, does not depend only on the accuracy of the classifications as “good” or “bad”, but also on the speed of the classification, the speed on which it can be revised and on the clarity of the model. According to these authors, ANNs are not good models due to their complexity and characteristic of “black boxes”; therefore, a model that is more intuitive and appealing is preferable, to clients and users, such as logistic regressions, probabilistic regressions and tree-based methods.

Alfaro and Gallardo (2012) conducted a study to assess what are the main determinants of consumer and mortgage default, at the household level in Chile, using data from the Survey of Household Finances made in 2007. The authors concluded that, at the consumer level, the main determinants are income-related variables, such as the number of people in the household that contribute with income; as well as the debt service ratio. At the

mortgage level, the authors also concluded that income-related variables are important, such as having a bank account and an education level beyond high school.

Besides the importance of sociodemographic and finance related variables (Abdou & Pointon, 2011; Avery, Calem, & Canner, 2004; Caouette et al., 1998; Constangioara, 2011; Costa, 2012; Gonçalves et al., 2013; Hand & Henley, 1997; Obrova, 2012) in order to construct a reliable model, it is also important to add variables that translate the change of the economic, health, or other conditions that may affect the ability of the client to pay back the money that was borrowed (Avery et al., 2004; Costa, 2012). This may be a health disease of some member of the family/household; a natural catastrophe, like the fires in Pedrogão Grande (Portugal), on June 2017; or some other unexpected *“economic or personal shock”* (Avery et al., 2004, p. 854). This is important because there are some circumstances that the client do not control and, therefore, are not related to its personal characteristics.

Just for curiosity, in 1982 some of the variables included in credit scoring models were if the household had a telephone at home and/or at the office, or not; the age difference between husband and wife; the zip code; and personal characteristics that nowadays are not allowed, like race, religion, sex, marital status and ethnic origin (Capon, 1982). This is relevant because highlights the importance of adapting the model as the years go by. With the development of technology, society, economy and with the emergence of new discoveries, the models must be adapted to translate the truth about individuals and their mutable behavior.

Table 3 presents the most used variables in the studies conducted by some authors. As it is possible to assess, variables related to sociodemographic information are the ones that appear the most, like income-related variables, age, marital status and level of education. On the other hand, despite the fact that they do not appear as much as sociodemographic information, variables related to the household's finances/credit history are also important, like having, or not, a bank account and a credit card, and having credit denied in the past.

	<i>Net wealth</i>	<i>Debt</i>	<i>Income</i>	<i>Age</i>	<i>Education</i>	<i>Time at current job</i>	<i>Personal sbocks</i>	<i>Number of dependents</i>	<i>Having credit denied</i>	<i>Home propriety</i>	<i>Regular expenses</i>	<i>Job situation</i>	<i>Mortgages</i>	<i>Gender</i>	<i>Marital status</i>	<i>Bank account</i>	<i>Time at current address</i>	<i>Home postcode</i>	<i>Type of credit</i>	<i>Occupation</i>	<i>Time at last job</i>	<i>Loan tenure</i>	<i>Loan period</i>	<i>Credit history</i>
<i>Greene (1992)</i>			x	x		x		x		x	x	x				x	x		x	x				
<i>Hand and Henley (1997)</i>			x	x		x				x					x	x	x	x	x	x				
<i>Constangioara (2011)</i>			x	x	x			x							x					x				x
<i>Alfaro and Gallardo (2012)</i>		x	x	x	x									x	x	x								
<i>Costa (2012)</i>		x	x	x	x		x	x			x	x	x											
<i>Samreen and Zaidi (2012)</i>			x	x		x		x							x					x	x	x	x	x
<i>Gonçalves et al. (2013)</i>			x	x		x								x	x		x	x	x	x				
<i>Henriques (2014)</i>	x	x	x	x	x	x	x	x	x	x														
<i>Absolute frequency</i>	1	3	8	8	4	5	2	5	1	3	2	2	1	2	5	3	3	2	3	5	1	1	1	2
<i>Relative frequency</i>	12.5%	37.5%	100%	100%	50%	62.5%	25%	62.5%	12.5%	37.5%	25%	25%	12.5%	25%	62.5%	37.5%	37.5%	25%	7.5%	62.5%	12.5%	12.5%	12.5%	25%

Table 3: Main variables included in the models of the mentioned authors

Chapter 3:

Data Description & Methodology

This chapter intends to describe the data used to develop this study as well as the methodology followed to pursue it, including basic statistical techniques and more advanced hypothesis tests. The chapter concludes with the description of the 27 variables that passed the different tests conducted and are, therefore, suitable for the development of the model that this study is trying to develop.

Part A: The Survey

The data used to develop this study was retrieved from a survey conducted by the European Central Bank in conjunction with the central banks of the Euro system and three National Statistical Institutes, in 2013. The survey, entitled “*Household Financial and Consumption Survey*” – HFCS –, provides detailed information on various aspects of European households, namely sociodemographic and financial information. The main questions of the survey are related to the property of the households inquired, like financial and fixed assets possessed; to possible loans that use those assets as collateral; as well as other financial obligations and applications. The survey includes, also, questions regarding heritages, income and the households’ decisions about consumption and savings, and questions regarding the individuals that compose the household, like age, level of education, and situation at the job.

The survey is a decentralized one. Each country that have contributed to the development of the survey worked individually and independently on their country. The Portuguese contribution was conducted by *Banco de Portugal* in conjunction with the Portuguese National Statistic Institute, which was one of the three National Statistics Institutes involved. The survey made by Portuguese entities is entitled “*Inquérito à Situação Financeira das Famílias*” – ISFF –, and it was conducted two times, one in 2010 and the other in 2013. The ISFF is composed by the same questions of the HFCS (designated as core

variables) as well as some questions oriented to the Portuguese type of families only. The 2013 survey inquired 8,000 Portuguese families that have resulted in 6,207 final households.

The Portuguese contribution is composed by more than 700 different variables, in which part of them concerns the household as a whole and the some concern each individual that composes each household, resulting in more than 16,000,000 observations, separated in 5 different files.

Part B: Methodology

As it was mentioned in the last chapter, in the past, credit institutions and credit analysts used their knowledge and prior experience when assessing the probability of default of some client. Later, that technique was systematized into the 5 C's of credit.

The authors in the literature defined 5 C's of credit: *character, capital, capacity, collateral* and *cycle conditions*. This study will consider the 5 C's of credit to assign an initial economic intuition behind the variables to include in the model, since it helps and have helped professionals assessing the likelihood of default of some clients. This study will assume the same number of C's but will substitute the first one – character – by a wider one – *personal and socio-professional characteristics* -, which is composed by three main sub-categories: “*Personal characteristics & Educational background*”; “*Professional & Financial situation*” and “*Family situation*”. This first “C” is important because gives a sense of the household's character and stability that is very important to predict if the household may default or not. Moreover, it gives a sense of the number of people in the household and if they are contributing to the household main income. The second “C”, capital, is important because gives an idea of possible resources available to use if an undesirable situation happens, considering more liquid assets, like financial assets, that are not used on a daily basis for regular expenses. Capacity concerns the volatility of earnings and the ability of the household to repay its debts, like, for example, the variable “income”. Collateral is also an important category because gives the notion if the household have assets that may be set as collateral. At last, the cycle conditions are important because they influence everyone and may have a very negative impact, even if a household is very wealthy.

This division helps the choice of the variables to include in the model, since this study relies on the ISFF, which has more than 700 different variables. Of those 700 variables, 68 were first selected, having in mind the ones used by the authors mentioned (see Table 3) as well as some that appeared to be relevant, due to an economic intuition. All of them are presented in Annex 2, grouped by category.

These 68 variables were first divided in two groups – continuous and categorical variables – since the analysis is different for each category of variables; and then each group was also divided in two different groups: households that have defaulted in the past and households that have not. This separation is important to assess if there is any significant difference among households who have defaulted and who have not, by looking at their variances, means, proportions, etc. The objective of this separation is to see which variables have informative content that enables the differentiation of households who have defaulted in the past and households who have not.

The first thing to do was to see if the distributions of the continuous variables had outliers and, if so, the second step was to control them using a technique called winsorization. This step is important because it will be necessary to calculate the mean and variance of the default and non-default groups to see if it is possible to differentiate among them. So, it is crucial to control for outliers because, in a distribution that is heavily skewed, the sample mean may not be the best estimate, since the difference between two sample means may offer a poor summary of how the populations differ and the magnitude of that difference (Everitt, 1992). To do that, the simple interquartile range statistic technique was used:

- First it is calculated the 1st (1Q) and 3rd quartile (3Q);
- Then it is calculated the interquartile range (IQR), which is equal to the difference between the 3rd and the 1st quartile;
- Finally, a sample has upper outliers if:

$$3Q + 1.5 * IQR < \text{Sample Maximum}$$

And bottom outliers if:

$$1Q - 1.5 * IQR > \text{Sample Minimum};$$

and the winsorization of the distribution was done using the software *E-views*.

Basically, to “winsorize” a distribution is to give less weight to values in the tails of the distribution and to pay more attention on those near the center, by substituting the highest x% of scores to the next smallest score and to change the x% smallest score to the next largest score (Everitt, 1992). To winsorize a distribution is, in a certain way, better than to “trim” a distribution - which is to simply delete the x% largest and smallest scores -, because using the winsorization technique no observations are lost.

Then, the variables that had outliers were winsorized at a 1% level and, if they still had outliers, they were winsorized at a 2% level, and so on. The variable with the higher

percentage of winsorization was “Total Financial Assets”, with a winsorization level of 12%, as Annex 3 shows.

After this, the first test computed was the Chi-square independence test, to both variables (continuous and categorical ones), to test if there is independence between an independent variable and the dependent one. The idea is to exclude variables that do not have a significant association with the dependent variable. It is important to note, though, that the relationship that this test tries to capture is not necessarily causal: one variable does not “cause” the other. The test is the following⁸:

Test 1. Chi-square test of Independence

H_0 : Variable A and variable B are independent

H_1 : Variable A and variable B are not independent

$$X^2 = \sum \frac{O_{r,c} - E_{r,c}}{E_{r,c}} \sim X^2(df)$$

where $O_{r,c}$ is the observed number of observations in row r and column c of the contingency table; $E_{r,c}$ is the number of estimated observations in row r and column c of the contingency table; r is the number of levels for one categorical variable; and c is the number of levels for the other categorical variable. The number of degrees (DF) of freedom is equal to:

$$DF = (r - 1) * (c - 1)$$

and the expected frequencies are computed separately for each categorical variable at each level of the other categorical variable, using the following formula:

$$E_{r,c} = \frac{n_r * n_c}{n}$$

where n_r is the total number of sample observations at level r of variable A , n_c is the total number of sample observations at level c of variable B , and n is the total sample size.

To compute this test, it was constructed a contingency table for each variable to help perform it and interpret the results⁹. The variables whose test concluded that are independent from the dependent variable were automatically excluded¹⁰. The ones that have passed the test are the following:

⁸ Every test was computed using a significance level equal to 5%.

⁹ In order to conduct this test among the continuous variables, they were divided into different classes in order to make this test feasible.

¹⁰ Since the significance level is considered to be 5%, the variables that presented a p-value higher than 5% were considered to be independent from the dependent variable and, therefore, automatically excluded.

<i>Nº</i>	<i>Variable</i>	<i>Chi-square</i>	<i>Degrees of freedom</i>	<i>P-value</i>
1	<i>Age</i>	8.66	3	3.41%
2	<i>Level of education</i>	63.58	1	0.00%
3	<i>Marital Status</i>	8.87	1	0.29%
4	<i>Level of education of the father</i>	10.96	2	0.42%
5	<i>Level of education of the mother</i>	10.19	2	0.61%
6	<i>Time at current job</i>	23.92	4	0.01%
7	<i>Credit denied</i>	13.08	1	0.03%
8	<i>Having a bank account</i>	40.16	1	0.00%
9	<i>Having credit card</i>	40.16	1	0.00%
10	<i>Having savings</i>	175.7	1	0.00
11	<i>Situation at current job</i>	76	3	0.00%
12	<i>Type of contract at current job</i>	11.85	1	0.00%
13	<i>Having another job</i>	4.86	1	2.74%
14	<i>Number of dependents</i>	20.60	3	0.01%
15	<i>Number of people in the household</i>	38.22	5	0.00%
16	<i>Number of people in the household with a job</i>	54.85	3	0.00%
17	<i>Type of residence</i>	8.49	3	3.68%
18	<i>Total residence surface</i>	39.94	4	0.00%
19	<i>Occupancy scheme</i>	32.66	3	0.00%
20	<i>Financial Assets</i>	212.77	1	0.00%
21	<i>Income</i>	137.81	4	0.00%
22	<i>Total expenses</i>	57.54	4	0.00%
23	<i>Expenses of the last 12 months in relation to income</i>	246.28	2	0.00%
24	<i>Expenses of the last 12 months in relation to the average</i>	11.03	2	0.40%
25	<i>Capacity to get financial support by friends and family</i>	8.61	1	0.33%
26	<i>Fixed Assets</i>	91.59	4	0.00%
27	<i>Wealth</i>	187.94	4	0.00%

28	<i>Having conditions deteriorated in the past 3 years</i>	55.47	1	0.00%
29	<i>Sector of the company where it has the main job</i>	16	8	3.68%
30	<i>Having conditions deteriorated in next 2 years</i>	5.74	1	1.66%
31	<i>Year of the acquisition of the main residence</i>	5.74	1	1.66%

Table 4: Variables that have survived Test 1

After that, the variances and means for the defaulted and non-defaulted groups were computed:

Test 2. Variance difference test

Since there are two tests to compute the test for the difference between two sample means, one assuming equal variances and one assuming the opposite, a test to infer the equality of the variances was necessary, which assumes the following hypothesis:

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

$$F = \frac{s_1^2}{s_2^2} \sim F(n_1 - 1; n_2 - 1),$$

where n_1 and n_2 are the number of observations of sample 1 and sample 2, respectively; and it is assumed that the population is normally distributed.

If the variances are assumed to be statistically different, Test 3.1 was computed, otherwise Test 3.2 was computed:

Test 3.1 Mean difference test with different variances

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{(\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2})^{1/2}} \sim t(n_1 + n_2 - 2)$$

where s_p^2 is the weighted average of sample variances and it is calculated as follows:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Test 3.2 Mean difference test with equal variances

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^{1/2}} \sim t(df)$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{1}}$$

Having in mind the last variable selection made using the first test, the continuous variables that have survived both tests are the following:

<i>N°</i>	<i>Variable</i>	<i>Mean</i>	<i>Mean Default</i>	<i>Mean Non-default</i>	<i>P-value (Test 2)</i>	<i>P-value (Test 3.1)</i>	<i>P-value (Test 3.2)</i>
1	<i>Age</i>	56 years	48 years	50 years	0.54%	1.31%	**
2	<i>Time at current job</i>	7.97 years	8.34 years	10.90 years	1.50%	0.00%	**
3	<i>Total residence surface</i>	129m2	124m2	137m2	41.89%	*	0.00%
4	<i>Financial Assets</i>	€18,421	€8,101	€19,060	0.00%	0.00%	**
5	<i>Income</i>	€110,623	€86,596	€136,675	0.00%	0.00%	**
6	<i>Total expenses</i>	€1,256	€1,243	€1,507	0.00%	0.00%	**
7	<i>Fixed Assets</i>	€224,531	169,214€	€257,000	0.00%	0.00%	**
8	<i>Wealth</i>	€160,203	€96,854	€160,564	0.59%	0.00%	**

Table 5: Variables that have survived Test 3.1 and Test 3.2

*: Test 3.2 was computed given the acceptance of the null hypothesis on Test 2.

**: Test 3.1 was computed given the rejection of the null hypothesis on Test 2.

As it can be seen when analyzing Table 4 and 5, all the continuous variables present in Table 5 are present in Table 4, which means that the difference in means test did not exclude any additional variable.

For the binomial categorical variables only, it was computed a binomial proportion test to assess if the difference in the proportions of observations “1” were statistically different, or not, from the ones equal to “0”. This test was made to exclude variables which

proportion is not significantly different in both groups (default and non-default groups) and can only be used as an exclusion test to the binomial variables. To the categorical variables that are not binomial, the test was useful to study the behavior of the dependent variable for each class of independent variable. The exclusion test to the binomial variables was the following:

Test 4. Two binomial proportions difference test

$$H_0: p_A - p_B = p_0$$

$$Z = \frac{\left(\frac{X_A}{N_A} - \frac{X_B}{N_B}\right) - p_0}{\sqrt{\frac{X_A(N_A - X_A)}{N_A^3} + \frac{X_B(N_B - X_B)}{N_B^3}}} \sim N(0,1)$$

where X_A/N_A and X_B/N_B represents the observed proportion in sample A and B, respectively.

Table 6 shows the results:

N^o	Variable	Z-test	P-value
1	Level of education	11.01	0.00%
2	Marital Status	2.86	0.21%
3	Credit denied	2.95	0.16%
4	Having a bank account	2.93	0.17%
5	Having credit card	6.18	0.00%
6	Having savings	16.34	0.00%
7	Type of contract at current job	2.80	0.26%
8	Having another job	2.74	0.31%
9	Capacity to get financial support by friends and family	2.81	0.25%
10	Having conditions deteriorated in the past 3 years	7.89	0.00%
11	Having conditions deteriorated in next 2 years	2.34%	0.96%

Table 6: Binomial variables that have survived Test 4

Once again, it is possible to conclude by analyzing both Table 4 and 6 that the test of the difference in proportions did not exclude any additional variable. Despite that, the test is important to make sure that the variables included in the model are, in fact, relevant for the purpose in question.

After the computation of all these different tests, the number of possible variables to include in the model was reduced from the initial 68 to a more reduced number, but yet a big one, equal to 32.

The following step was to see if there is correlation between those variables, in each category, in order to avoid future collinearity problems. By constructing the correlation matrix, it was possible to conclude the immediate exclusion of some variables and the inclusion of new transformed variables. For example, it was included the ratio between people in the household with a job and the number of people in the household, instead of having these two variables separately; and a variable including the higher degree of education of the parents of the representative person of the household, instead of having, again, the two variables separately. By doing this transformation, the number of possible variables went from 32 to 27, with no high correlation among them (lower or equal to $|50\%|$).

The resulting variables are the following:

1. Personal and Socio-professional characteristics

1.1 Personal characteristics & educational background

1. *Age;*
2. *Level of education;*
3. *Marital status;*
4. *Higher level of education obtained by the parents of the household's representative;*

1.2 Professional & financial situation

5. *Time at current job;*
6. *Credit denied;*
7. *Having a credit card;*
8. *Having savings;*
9. *Situation at the current job;*
10. *Having another job;*
11. *Type of contract at current job;*

1.3 Family situation

12. *Number of dependents;*
13. *Ratio between the number of people in the household with a job and the number of people in the household;*
14. *Type of residence;*
15. *Total residence surface;*
16. *Occupancy scheme;*

2. Capital

17. *Total Financial Assets*

3. Capacity

18. *Expenses of the last 12 months in relation to income;*
19. *Expenses of the last 12 months in relation to the average;*
20. *Capacity to get financial support by friends and family;*
21. *Income;*
22. *Regular Expenses/Income;*

4. Collateral

23. *Wealth (without financial assets);*

5. Cycle Conditions

24. *Having conditions deteriorated in the past 3 years;*

25. *Sector of the company where it has main job;*

26. *Having conditions deteriorated in the next 2 years;*

27. *Year of acquisition of the main residence.*

All variables that were tested, excluded or created throughout this process are present in Annex 4, together with their respective tests results.

Part C: Data Description

This subsection of the chapter intends to describe the set of variables that have survived the upper-mentioned tests and, consequently, may incorporate the final model that this study is trying to develop.

The dependent variable reflects the default or delay in the payments of the last 12 months. This variable is composed by 3,398 observations of a total of 6,207 inquired families. 12% of the households have defaulted or delayed in the payments in 2012 (12 months before 2013) and 88% have not.

The independent variables, as it was already mentioned, are divided in 5 categories, where the qualitative ones were transformed into *dummy* variables.

The independent variables are the following:

1. Personal and Socio-professional characteristics

1.1 Personal characteristics & Educational background

1. *Age:*

The variable age concerns the person responsible for the household and it is measured in years. Being a quantitative variable, its mean is equal to 56 years. As it was expected, the minimum age of the representative person of the household is equal to 18 years old, and the maximum is equal to 90 years.

As it is possible to conclude by looking at Table 3, this variable is commonly used in credit scoring models, since it was used by Greene (1992), Hand and Henley (1997), Constangioara (2011), Alfaro and Gallardo (2012), Costa (2012), Samreen and Zaidi (2012), Gonçalves et al. (2013), and Henriques (2014), which represents 100% of the mentioned authors.

2. Level of education

This variable also concerns the responsible for the household and it was arranged in order to present two possible outcomes: 1, if the representative has superior education (bachelor's degree or above); and 0, if not. Of those 6,207 household's representatives inquired, only, approximately, 20% have superior education. This variable may be important because it is believed that the level of education may give a superior decision-making ability and, therefore, may diminish the number of defaults; moreover, it is expected to see people with superior education having better salaries, representing less default probabilities.

This variable is also commonly used in the literature and it was used by 50% of the mentioned authors, such as Constangioara (2011), Alfaro and Gallardo (2012), Costa (2012) and Henriques (204).

3. Marital status

This variable also concerns the representative of the household and may be important because, usually, people that are married tend to pay more regularly their debts. This happens because the probability of having two incomes contributing to the household's main income is higher and, therefore, the probability of default decreases.

Of those 6,207 households inquired, 65% are married and 35% are not.

This variable was used by some of the authors mentioned, namely Hand and Henley (1997), Constangioara (2011), Alfaro and Gallardo (2012), Samreen and Zaidi (2012) and Gonçalves et al. (2013).

4. Higher level of education obtained by the parents of the household's representative

The variable "higher level of education obtained by the parents of the household's representative" was constructed with information regarding the level of education of the parents of the household's representative, present in the ISFF. This variable and the dependent variable are not independent, as the chi-square test of independence suggests, and may be relevant to the model since not everyone has the same background and people with parents with financial capacities may have financial capacities to honor their debt, even if they do not have a good job/income.

The level of education of the parents of the household's representative may present three different outcomes: basic education (lower than high school), high school, or superior education (higher than high school). Among the 6,102 households, only 6.96% of the parents

have superior education (at least one of them), 5.11% completed high school and 87.92% have basic education.

To include this variable in the credit scoring model, it was transformed into 2 dummy variables.

1.2 Professional & financial situation

5. Time at current job

This variable concerns the household's representative, being expected to see more financial and professional stability as the years in the same job or company increases. Since it can only be answered by people with a job, and in order to not lose a huge amount of observations, people that are unemployed, domestic, retired, are studying or are disable or inactive were considered to be in the "company" for 0 years. Having this in mind, the number of observations considered were 6,191 households (instead of only 3,209 if only workers were considered), with mean, minimum and maximum equal to 8, 0 and 55 years, respectively. Since the distribution of these variables has outliers, they were controlled to compute the tests, to avoid biases in the results.

This variable was also considered by Greene (1992), Hand and Henley (1997), Samreen and Zaidi (2012), Gonçalves et al. (2013) and Henriques (2014).

6. Credit denied

This variable, as the title suggests, translates if a household have had credit denied in the past, or not. It does not only concern the household's representative but everyone in the household that have asked for credit in the last 12 months (therefore, in 2012). This variable may be important because situations that have happened in the past may be repeated in the future, since may reflect the person's character and way of living.

Giving that not every household have applied to credit in the last 12 months, this variable is composed by only 951 observations, from which 819 (86.12%) have not been denied credit in the past, and the others 132 (13.88%) have been denied credit in the past 12 months.

This variable was also considered by Henriques (2014), since we are both inspired by the ISFF answers; and by Constancioara (2011) and Samreen and Zaidi (2012), when considering the credit history of the client.

7. Having a credit card

At first, the variable that was supposed to be considered was "having a bank account" but since most households have a bank account nowadays, the variable considered was if the

household has a credit card, or not. Of the 6,207 households inquired, 45.90% responded that they do not own a credit card, and the rest (54.10%) own a credit card.

8. *Having savings*

This variable translates if the household possesses any kind of savings, by having a savings account at the bank or savings of any kind. Of those 6,207 households inquired, 3,168 (51.04%) have responded that they possess savings, and the other 3,039 (48.96%) do not. Of the group of households that have defaulted in the past, only 18.87% of the households have savings, which is supported by the chi-square test of independence, which rejects the null hypothesis that states that these two variables – having savings at the bank or at home and the dependent variable – are statistically independent.

9. *Situation at the current job*

This variable has the objective to clarify what is the situation at the job of the household's representative. That is, if the representative is (i) a regular paid worker; (ii) a worker on leave; (iii) unemployed; (iv) a student; (v) retired; (vi) disabled; (vii) domestic; or (viii) other inactive. Annex 5 shows the number of households in each situation.

As it was expected, the biggest pie belongs to the regular paid workers that represents 51.64% of the households, while the smallest pie belongs to the students (0.19%). To incorporate this variable in the model, it was transformed into 7 dummy variables, corresponding to 8-1 categories.

This variable was also included in the studies of Greene (1992) and Costa (2012).

10. *Type of contract at current job*

The variable “type of contract at current job” may be a very good indicator of the behavior of people depending on their type of contract. This variable is a binary one and was constructed to present two possible outcomes: 1, if the contract has a maturity; and 0, if not. At first glance, it is expected to see a more concerned behavior among people that have contracts with maturity, because their professional future and, therefore, their future income, are not as guaranteed as the ones by those households that do not have maturities in their contracts.

From a total of 1,894 observations, and concerning only the household's representative, 197 households' representatives have maturities in their contracts, while the others 1,697 do not.

11. Having another job

This variable may be a very relevant variable to include in the model because may reflect an additional ability to pay the debts, since having another job indicates another source of income. This variable includes 3,351 observations, of which 256 household's representatives indicate having another job; while the others 3,095 do not. Comparing the percentage of households that have more than one job between the households that have defaulted in the past and the ones that did not, 4.89% and 8.87% of them have another job, respectively. This difference in the proportions seems to be statistically relevant, as the difference in proportions test defends.

1.4 Family situation

12. Number of dependents

The number of dependents concerns the entire household and is composed by the children (people with 18 or less years old) that composes the household. This variable is important because it is expected to see a higher probability of default among households that have more dependents since they bring expenses and, usually, do not contribute with income.

In a total of 6,207 observations, 4,014 households have zero children with 18 or less years; 1,251 have one; 768 have two; 145 have three; and 29 have 4 or more.

This variable, or one closely related, was also included in the models of Greene (1992), Constangioara (2011), Costa (2012), Samreen and Zaidi (2012) and Henriques (2014).

13. Ratio between the number of people in the household with a job and the number of people in the household

Just like the variable “higher level of education obtained by the parents of the household's representative”, this variable was created after the computation of the correlation matrix. Since the variables “number of people in the household” and “number of people in the household with a job” were correlated ($\rho = 50.06\%$), their ratio was computed to avoid collinearity problems.

This variable is composed by 6,207 observations and has mean, minimum and maximum equals to 36.67%, 0% and 100%, respectively. This 36.67% means that, in a household composed by 5 people, approximately two (1.83) of them have a job.

14. Type of residence

This variable concerns the type of the residence where the household lives. This variable is a categorical one and may have 4 outcomes: (i) apartment; (ii) individual habitation; (iii) townhouse; or (iv) other. 3,156 households live in an apartment, 3,050 households live

in a townhouse; while 1,693 households live in an individual habitation; and the other 51 live in another kind of habitation.

After further analysis, this variable had to be excluded from the list of possible variables due to the way it was constructed. Since this variable was responded by the interrogator of the survey by observing the house of the household, it contains 8,000 observations instead of 6,207, making it impossible to include in the model due to misalignment of observations.

15. Total residence surface

This variable, as total financial or non-financial assets, may tell a little about the household's wealth, through the number of m² of its main residence. Just like all the other variables presented in this study, it was retrieved from the ISFF 2013, where it is possible to assess that the average surface of the Portuguese's household's houses is equal to 129.22m², with 10m² as a minimum and 200m² as a maximum. As expected, it is measured in m².

16. Occupancy scheme

The variable "occupancy scheme" tries to capture the type of occupancy that the different households possess. As it was possible to conclude, in a total of 6,207 households, 4,898 of them have the total ownership of the house where they live in; 155 have co-ownership; 781 rent their houses; and the remaining 373 live in their houses for free.

Being a categorical variable with more than two different outcomes, it was transformed into three dummy variables, to make it possible to include in the final model.

2. Capital

17. Total Financial Assets

Even though this variable could have been incorporated in the variable "wealth" for, in fact, being a part of the household's wealth, it was separated from it for being a more liquid type of wealth, readily available – or at least more readily available – when needed.

These financial assets include the following assets: current accounts; savings accounts; investment funds; treasury bonds; investments in a company; shares; accounts managed by clients' manager; other assets; value of credit conceded to friends and family; other financial assets; and mutual funds.

According to the survey, the average amount of financial assets possessed by the Portuguese families in 2013 was, approximately, €32,107; the minimum was €0.00, and the maximum was €2,740,182. By calculating the interquartile range, it was possible to infer the presence of outliers that were then controlled with winsorization at a 12% level in the upper tail for the computation of the tests mentioned (as Annex 3 shows).

3. Capacity

18. Expenses of the last 12 months in relation to income

This variable was responded by the household's representative and represents the relation between the regular expenses of the household of the last 12 months and the income received by the household during the same period of time. The variable may take three different outcomes (that were converted into two dummy variables): regular expenses superior than the income; inferior; or similar. The first dummy variable constructed takes the value 1, if the expenses are superior than the income; and 0, otherwise. The second dummy variable takes the value 1, if the expenses are inferior than the income; and 0, otherwise.

From a total of 6,199 families inquired, 915 responded that the regular expenses of the last 12 months were superior than the income of the total household; 2,197 responded that the expenses were inferior than the income; and the remaining families – 3,087 - responded that the expenses and the income of the last 12 months were more or less similar.

19. Expenses of the last 12 months in relation to the average

This variable concerns the entire household, representing the relation of the regular expenses supported by the household and the expenses that are assumed to be the average. The question made was: *“Do you consider the regular expenses of the last 12 months to be superior, similar or inferior than the regular expenses of a normal year?”*.

2,138 families responded that the expenses were superior; 735 responded inferior; and the remaining 3,323 responded similar.

To incorporate this variable in the credit-scoring model, it was transformed into two dummy variables. Just like in the variable 18, the first dummy variable takes the value 1, if the expenses are superior than the average; and 0, otherwise. The second dummy variable takes the value 1, if the expenses are inferior than the average; and 0, otherwise.

20. Capacity to get financial support by friends and family

This variable is related to the capacity of getting financial support by friends and family anytime the household needs, and it refers to the whole household. This variable was transformed into a dummy variable that takes the outcome 1, if the household has the capacity to get financial support from friends and/or family; and 0, otherwise. If the outcome is equal to 1, it means that the household is able to get money when they need to pay some debt or when an undesirable and unexpected situation happens. Obviously, it is a subjective variable that may not be really true to reality.

In the ISFF, this variable includes 6,145 answers, where the majority ($\approx 70\%$) states having the capacity to get financial support by friends and family.

21. Income

The variable income concerns the whole household and it is the sum of the value of the different incomes that the household receives, namely: employee income; self-employment income; income from pensions (income from public, occupational and private pension plans); regular social transfers (except pensions); income from regular private transfers; gross rental income from real estate proprietary; gross income from financial investments; gross income from private businesses other than self-employment; and residual income variable.

This variable is measured in euros and is composed by 6,207 observations, with mean, minimum and maximum equal to €119,118.13; €0.00; and €3,802,500.00, respectively. As Annex 3 demonstrates, the distribution of this variable contained outliers that were controlled with a winsorization level equal to 6%. The new mean, minimum and maximum are equal to €110,623.00; €0.00; and €300,000.00, respectively.

As it can be seen by looking at Table 3, this variable is commonly used in the studies of the authors mentioned, namely Greene (1992), Hand and Henley (1997), Constangioara (2011), Alfaro and Gallardo (2012), Costa (2012), Samreen and Zaidi (2012), Gonçalves et al. (2013) and Henriques (2014).

22. Regular Expenses/Income

This is the third variable created after the computation of the correlation matrix. As it was expected, regular expenses and income are heavily correlated due to the fact that, usually, who earns more money also tends to spend a little more as well. Thanks to that, we find convenient to create a ratio that compares one with the other.

So, this ratio is created by dividing the annual regular expenses by the annual gross income received by the whole household; and then multiplied by 100 to get the percentage. The minimum percentage of this ratio is 0.12%, which means that the household spends only a small part of its income, because they spend very little or because their income is more than enough to cover the expenses; while the maximum value was an odd one equal to 23,157.89%, meaning that the household spends way more than they earn. The mean value was a normal one, equal to 36.67%. Having a huge value for the maximum, this variable had obviously outliers that were controlled for the computation of the tests. After outliers were controlled, the average of this variable was equal to 16.59% (see Annex 3).

From a total of 6,144 observations, 340 of them have ratios superior than the mean (36.67%) and 84 are higher than 100%, which means that 84 households spend more money than they earn.

4. Collateral

23. Wealth (without financial assets)

The value of wealth refers to the whole household and it is calculated as the difference between the value of the assets possessed by the household (excluding financial assets) and the value of liabilities and other responsibilities. To construct this variable, the assets considered were “current residence value”; “current value of other residences”; “current value of what owns”; “current value of automobiles”; “current value of other vehicles”; “current value of high value objects”; and “net value of participation in a company”.

The wealth is measured in euros and it includes 6,207 observations. Its mean is approximately €222,000, with minimum and maximum values equal to €-207,500 and €20,747,892, respectively. As it was expected, this variable includes observations that are considered to be outliers that were properly controlled for the computation of the tests needed (see Annex 3).

This variable was also considered by Henriques (2014); and some authors included the variable debt in their models, namely Alfaro and Gallardo (2012) and Costa (2012), but in an isolated way. This study includes the variable debt incorporated in the wealth of the households due to its higher explanatory power and the fact that the variable debt alone was statistically independent of the dependent variable, which would not contribute with anything in the estimation of the probability of default.

5. Cycle Conditions

24. Having conditions deteriorated in the past 3 years

This variable concerns the entire household and may take two possible outcomes: 1, if a member of the household has seen his/her conditions deteriorated in the last 3 years (prior to the survey that was conducted in 2013); and 0, otherwise. Having conditions deteriorated means that some member (i) have lost his/her job; (ii) have had to work fewer hours; (iii) have had to accept non-desirable changes at the job; or (iv) other.

Of those 6,207 households inquired, 2,469 state having conditions deteriorated in the past 3 years, while 3,738 don't.

This variable was included in the models of Costa (2012) and Henriques (2014), who are authors that have relied in the ISFF of 2010.

25. Sector of the company where it has main job

This variable concerns the representative of the household and may have a significant impact on the probability of default of a household because it tells a little about the conditions that the representative faces. As we all know, the cycle conditions of the economy do not affect all sectors of the economy at the same time, so, the sector of the company may predict, at some point, the default.

This variable is a categorical one and, as Annex 6 shows, may have 12 different outcomes: (i) Agriculture, animal production, hunting, forest and fishing; (ii) Extractive and transformative industries, electricity, gas, steam, water, ..., waste management and decontamination; (iii) Construction; (iv) Wholesale, retail and vehicle repair; (v) Transportation and storage; (vi) Accommodation and catering; (vii) Communication and information services; (viii) Finance and insurance services; (ix) Public and defense administration; (x) Education; (xi) Health and social support; and (xii) Artistic Activities.

To make it possible to include this categorical variable in the model, it was transformed into 11 dummy variables.

26. Having conditions deteriorated in the next 2 years

This variable, just like the variable “Having conditions deteriorated in the past 3 years”, concerns the entire household and may take two possible outcomes: 1, if the household expects to see his/her conditions deteriorated in the next 2 years; and 0, otherwise. If the outcome is equal to 1, it means that, in the next 2 years, one member of the household expects to (i) lose his/her job; (ii) work fewer hours; (iii) accept non-desirable conditions at the work; or (iv) other.

Of those 3,273 households that were inquired, 1,207 of them expects to see their conditions deteriorated in the near future; while 2,066 don't.

27. Year of acquisition of the main residence

At last, the final variable is also related to the cycle conditions of the economy and it is the year of the acquisition of the main residence of the household.

The output of the variable may be any year since 1940 until 2013. As it can be seen in Annex 7, most houses were acquired between 2001 and 2010, representing more than half of the houses acquired in the period of observation.

Chapter 4:

The Model

Part A: The Model

After the computation of the tests mentioned in the last chapter, the model creation is now feasible. To do so, the second stage of the variable's selection went through the choice of the best combination of the variables, including the ones with the higher significance power, and variables from each of the 5 categories previously defined: personal and socio-professional characteristics; capital; capacity; collateral; and cycle conditions.

The type of model chosen was the probabilistic one (probit model), because it is used by some authors in the literature (like Greene (1992), Alfaro and Gallardo (2012) and Henriques (2014)); it is the one that presents the best results in terms of accuracy, as Table 2 shows; it is easy to understand and implement, being very intuitive; and it was the one used by Henriques (2014), which this study is trying to update and improve. This kind of regression is indicated when the dependent variable (Y) is assumed to be binary assuming only two possible outcomes. This technique aims to maximize the likelihood of an event happening, translating into better estimates of the coefficients of the explanatory variables.

In this specific case, the variable Y reflects the default or delay in the payments of the last 12 months, and it takes the value 1, if the household has defaulted or delayed in any payment in the last 12 months; and 0, otherwise. This model assumes that the binary outcomes are mutually exclusive, which means that one household either defaults or delays, or not. The outcome of the model is the probability of Y being equal to 1, which is the probability of default, given some attributes (independent variables, X):

$$P = P(Y = 1|X) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k),$$

where Φ is the normal cumulative distribution function.

The estimated model is defined as:

$$\begin{aligned}
PD_i = Probability(default)_i &= \Phi (\beta_0 + \beta_1 Age_i + \beta_2 Level\ of\ Education_i + \beta_3 Marital\ Status_i \\
&+ \beta_4 Time\ at\ current\ job_i + \beta_5 Credit\ denied_i + \beta_6 Having\ savings_i \\
&+ \beta_7 Number\ of\ dependents_i \\
&+ \beta_8 Occupancy\ scheme\ 1: Total\ ownership_i \\
&+ \beta_9 Occupancy\ scheme\ 2: Co - ownership_i \\
&+ \beta_{10} Occupancy\ scheme\ 3: Rent_i + \beta_{11} Total\ Financial\ Assets_i \\
&+ \beta_{12} Expenses\ of\ the\ last\ 12\ months\ in\ relation\ to\ income\ 1: Superior_i \\
&+ \beta_{13} Expenses\ of\ the\ last\ 12\ months\ in\ relation\ to\ income\ 1: Inferior_i \\
&+ \beta_{14} Income_i + \beta_{15} Wealth\ (without\ financial\ assets)_i \\
&+ \beta_{16} Having\ conditions\ deteriorated\ in\ the\ past\ 3\ years_i + \varepsilon_i)
\end{aligned}$$

where i is the family observed, β_0 is the constant of the model, β_1 until β_{16} are the coefficients of the explanatory variables and ε is the error of the model that is not explained by the included variables.

To make sure that the variables chosen are the best to the model, the tests mentioned in the last chapter were conducted prior to the creation of the models (variance and mean difference test, proportion test and the chi-square independence test); as well as the correlation matrix, to avoid collinearity problems.

The following table (Table 7) shows the coefficients and the standard errors of the model, when applied to the whole sample. The objective of this exercise is to infer the significance of the variables and if the signs of the variables are in accordance to what is expected. As it can be seen when analyzing Table 7, most variables are in accordance to what is expected. For example, as the number of dependents increase, it is expected to see a higher probability of default, since more children mean more expenses; and the negative sign of the variable “number of dependents” confirms that. The third column of the table shows if the variable is in accordance to what is expected – represented by the symbol “✓” –; if not – represented by the symbol “X”; or if it can take both signs – represented by both signs “✓X”.

Model 0

<i>Variable</i>	<i>Coefficient</i>	<i>Std. Error</i>	<i>Economic intuition</i>
<i>C</i>	-1.178743 ***	0.395274	-
<i>Age</i>	0.001442	0.005495	✓X
<i>Level of Education</i>	-0.536004 ***	0.183413	✓

<i>Marital Status</i>	-0.243814 *	0.133064	✓
<i>Time at current job</i>	-0.006285	0.006254	✓
<i>Credit denied</i>	0.326398 **	0.154826	✓
<i>Having savings</i>	-0.426015 ***	0.131997	✓
<i>Number of dependents</i>	0.167967 **	0.071667	✓
<i>Occupancy scheme 1: Total ownership</i>	0.050220	0.257458	✗
<i>Occupancy scheme 2: Co-ownership</i>	-0.273387	0.489190	✓✗
<i>Occupancy scheme 3: Rent</i>	0.004289	0.289349	✓
<i>Total financial assets</i>	2.31E-06 ***	8.79E-07	✗
<i>Expenses of the last 12 months in relation to income 1: Superior</i>	0.442413 ***	0.162451	✓
<i>Expenses of the last 12 months in relation to income 2: Inferior</i>	-0.253034	0.155906	✓
<i>Income</i>	-8.02E-07	8.77E-07	✓
<i>Wealth (without financial assets)</i>	-4.90E-07 *	2.73E-07	✓
<i>Having conditions deteriorated in the past 3 years</i>	0.545304 ***	0.131841	✓
<i>McFadden R-squared</i>	0.162733		
<i>Akaike info criterion</i>	0.689396		
<i>Total observations</i>	863		
<i>Observations with Dep=0</i>	750		
<i>Observations with Dep=1</i>	113		

Table 7: Model 0

*, *p*-value < 0.1

**, *p*-value < 0.05

***, *p*-value < 0.01

Then, to compute the model, the sample was divided in two groups with the same number of observations, randomly created, one for training and one for testing, just like Šušteršič et al. (2009) defend in their works.

Constructing the first model – model A – most observations were lost (model A contains only 428 observations), mainly due to the variable “credit denied” because not every household has asked for credit, so not every household had been denied credit; and since most variables seem not to be significant, a similar model without this variable was regressed – model B.

The variables of the models tested – model A and B - are defined as follows:

<i>Variable</i>	<i>Model A</i>		<i>Model B</i>	
	Coefficient	Std. Error	Coefficient	Std. Error
<i>C</i>	-0.857289	0.549710	-0.784886 ***	0.294125
<i>Age</i>	7.67E-05	0.007351	-0.001576	0.004022
<i>Level of Education</i>	-0.345102	0.246808	-0.466927 ***	0.137839
<i>Marital Status</i>	-0.327236 *	0.188156	-0.148022	0.098237
<i>Time at current job</i>	-0.000448	0.008512	-0.003341	0.004349
<i>Credit denied</i>	0.326195	0.228892	-	-
<i>Having savings</i>	-0.222721	0.181770	-0.497798 ***	0.097243
<i>Number of dependents</i>	0.149543	0.095101	0.210630 ***	0.049161
<i>Occupancy scheme 1: Total ownership</i>	-0.306222	0.363722	-0.252278	0.190904
<i>Occupancy scheme 2: Co-ownership</i>	-0.455037	0.577382	-0.079066	0.335560
<i>Occupancy scheme 3: Rent</i>	-0.160822	0.399470	-0.243722	0.221801
<i>Total financial assets</i>	1.57E-06	1.05E-06	7.13E-07	8.14E-07
<i>Expenses of the last 12 months in relation to income 1: Superior</i>	0.344007	0.222469	0.656140 ***	0.120695
<i>Expenses of the last 12 months in relation to income 2: Inferior</i>	-0.269169	0.218520	-0.057609	0.107674
<i>Income</i>	-1.21E-06	1.25E-06	-1.95E-06 ***	6.94E-07
<i>Wealth (without financial assets)</i>	-2.82E-07	3.30E-07	1.62E-08	1.07E-07
<i>Having conditions deteriorated in the past 3 years</i>	0.583149 ***	0.184904	0.422023 ***	0.090076
<i>McFadden R-squared</i>	0.140818		0.174014	
<i>Akaike info criterion</i>	0.768576		0.630024	

<i>Total observations</i>	428	1703
<i>Observations with Dep=0</i>	369	1496
<i>Observations with Dep=1</i>	59	207

Table 8: Model A and model B

*: $p\text{-value} < 0.1$

**: $p\text{-value} < 0.05$

***: $p\text{-value} < 0.01$

Looking at the signs of the coefficients, it is possible to conclude that most of them follow what is economically expected, in exception of the variables “Occupancy scheme: Rent” and “Total financial assets”, on model A; and the variables “Occupancy scheme: Rent”, “Total financial assets” and “Wealth” on model B, despite the different results when the model is applied to the whole data (model 0).

In a probabilistic or logistic model, that has a binary response, the best way to analyze the model’s performance is by looking at the accuracy rates computed with a predetermined cut-off. The accuracy rate is the percentage of clients that the model has successfully predicted as good or bad ones. For example, with a 50% cut-off, clients that have a probability of default equal or higher than 50% are considered to be bad clients – clients that won’t be granted credit because they have a high probability of default –; and clients with a probability of default smaller than 50% are conceded credit. Then, according to each probability obtained, each one of them is compared to the true reality – if the client has actually defaulted, or not, in the past 12 months –, and the accuracy rate is obtained by dividing the first one by the second. Usually, in the literature, the cut-off that is commonly used is 50%, since one “*should predict default if the model predicts that it is more likely than not*” (Greene, 1992, p. 6). In this study, however, different cut-offs are going to be computed since a cut-off equal to 50% is considered by Banco Carregosa to be too high, since its credit analysts tend to use cut-offs around 10%. By using different cut-offs, it is possible to infer which cut-off is the best; and the one that produces the higher total accuracy, having in mind that the accuracy rate that this study is most focused on is the accuracy rate of the defaulted clients. The reason behind it relies on the fact that banks prefer to avoid bad clients that won’t pay instead of losing clients that would have paid.

This study will present the outcomes of the model with a 50%, 30%, 20%, 15% and 10% cut-offs. Since the 15% cut-off is the one that presents higher accuracy rates, the other outcomes are present in Annex 8.

Model A		Observed		
Cut-off=15%		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 15\%$	39	98	137
	$<15\%$	15	283	298
	<i>Total</i>	54	381	435
Total accuracy		74.02%		
“Default Accuracy”		72.22%		
“Non-default accuracy”		54.28%		

Table 9: Accuracy of the model A with a 15% cut-off

Model B		Observed		
Cut-off=15%		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 15\%$	132	297	429
	$<15\%$	68	1184	1252
	<i>Total</i>	200	1481	1681
Total accuracy		78.29%		
“Default Accuracy”		66.00%		
“Non-default accuracy”		79.95%		

Table 10: Accuracy of the model B with a 15% cut-off

As it can be seen by analyzing the tables (Annex 8), the consideration of a cut-off different than 50% is a very crucial step in the creation of a model. It is possible to conclude that the cut-off that presents the higher accuracy rates is the cut-off equal to 15%. As it can be seen, with a 15% cut-off, the model A presents a total accuracy of 74.02% and an accuracy rate to the default group equal to 72.22%. This accuracy is a very good one, since the model is able to predict almost three quarters of the defaulted clients. On the other hand, model B seems to present better rates, since the total accuracy rate is higher – equal to 78.29%, despite the fact that the “default accuracy” is lower – equal to 66.00%. Moreover, model B may be more reliable due to the higher number of observations included; and to the higher number of variables that are significant.

In a practical point a view, the best model to use is the model B, the one without the variable “credit denied”. A possible reason for that relies on the fact that the people that ask for credit may have incentives to lie and tell that never had credit denied in the past in order to make a better impression and get the credit that they want. Since most banks can’t have this type of information confirmed, unless it was them that have denied credit to the client,

the better way to avoid such problem is to not include variables that the clients may have incentives to lie about.

To further compare these models with others in the literature and infer about their performance, the next sections will present 3 different models: the first one, developed by Henriques (2014), using the ISFF 2010; the second one is the same model, but estimated now; and the third one is the model proposed by Saunders and Cornett (2012) , a simple credit rating model. All these models were applied to the same data, which are the responses made to the ISFF 2013.

Part B: Comparison with other models

1. Henriques (2014)' Model – Version 1 and 2

The original model developed by Henriques (2014) is a probabilistic model, just like the one we presented in the previous section.

Her estimated model, regressed with information retrieved from the ISFF 2010, is defined as follows:

$$\begin{aligned}
 PD_i &= Probability(default)_i \\
 &= \Phi (\beta_0 + \beta_1 Net\ wealth_i + \beta_2 Income_i + \beta_3 Age_i + \beta_4 Level\ of\ education_i \\
 &\quad + \beta_5 Time\ at\ current\ job_i \\
 &\quad + \beta_6 Having\ conditions\ deteriorated\ in\ the\ past\ 3\ years_i \\
 &\quad + \beta_7 Number\ of\ children_i + \beta_8 Credit\ denied_i + \beta_9 Propriety\ ownership_i \\
 &\quad + \varepsilon_i)
 \end{aligned}$$

where i is the family observed, β_0 is the constant of the model, β_1 up until β_9 are the coefficients of the explanatory variables and ε is the error of the model that is not explained by the included variables.

In order to compare this model with the one presented in the previously and the following sections, Annex 9 presents the accuracy rates of the model, at different cut-offs, when applying the model to the ISFF 2013's data. The purpose of this exercise is to study the robustness of the model and its coefficients and the impact that the time has on them; as well as to compare it with the other models presented.

As it can be seen by analyzing the tables in Annex 9, the accuracy rate of the model is very acceptable, with an accuracy near the 60% when using a cut-off equal to 15%. The results are not as high as the ones from model A and B but, since there is a time gap between the creation of the model and the data applied, I consider it as a reasonable model.

Since Henriques (2014)' model was regressed using the ISFF of 2010, a new model was regressed, again, using the data of the ISFF 2013.

Just like the other version of Henriques (2014)' model, this new model follows a probabilistic regression, with the same expression as before. As it can be seen at Table 11, the variables included are the same as the previous model presented; the only thing that changes are their coefficients, since the model was regressed using different data. Just like the models presented in the last chapter, to compute this model, the sample was divided in two samples, with the same number of observations, one to develop the regression and the other to test it.

As Table 11 shows, the variables from both models – Henriques (2014)' Model version 1 and 2 – as well as their definition.

<i>Variable</i>	<i>Henriques (2014)'s Model – Version 1</i>		<i>Henriques (2014)'s Model – Version 2</i>	
	<i>Coefficient</i>	<i>Std. Error</i>	<i>Coefficient</i>	<i>Std. Error</i>
<i>C</i>	-1.5406 ***	0.5232	-1.250106 ***	0.467083
<i>Net wealth</i>	-1.7E-06 *	9.2E-07	-3.74E-07	4.11E-07
<i>Income</i>	3.1E-06	4.1E-06	-1.99E-07	1.14E-06
<i>Age</i>	0.0026	0.0098	-0.003251	0.007973
<i>Level of education</i>	-0.0279	0.2681	-0.387004	0.336148
<i>Having conditions deteriorated in the past 3 years</i>	0.5211 **	0.2133	0.552027 ***	0.174783
<i>Number of children</i>	0.2263 **	0.0980	0.073402	0.100679
<i>Credit denied</i>	1.1145 ***	0.2128	0.544418 ***	0.223433
<i>Time at current job</i>	-0.0228 **	0.0113	-0.014103	0.008804
<i>Propriety ownership</i>	-0.3248	0.2317	0.095969	0.208983
<i>McFadden R-squared</i>	0.2520		0.089691	
<i>Akaike info criterion</i>	0.26373		0.711702	
<i>Total observations</i>	364		428	
<i>Observations with Dep=0</i>	n/a		377	
<i>Observations with Dep=1</i>	n/a		51	

Table 11: Model developed by Catarina Henriques (2014)' model – version 1 and 2

*: p-value < 0.1

**: p-value < 0.05

***: p-value < 0.01

n/a: information not available.

Just like in the other models presented in the previous sections, in order to compare all of them, the results of version 2 of Henriques (2014)'s model, with different accuracy rates, were computed and are presented in Annex 10.

When looking at the table that presents the accuracy rates when using a cut-off equal to 15% (see Annex 10) and, comparing them with the Catarina Henriques (2014)'s original model, the results are more or less the same. The total accuracy of the models is very similar – 61.34% for the original model and 60.09% for the regressed one –, but the “default accuracy” is way better for the regressed one – 60.18% for the first versus 69.35% for the second. These better results show that it is better to regress a new expression as soon as new data is available, since the data is more recent and translate better the new aspects of the economy and the clients. Having this in mind, the incorporation of this model in this study is important to compare this model with the models that I proposed – model A and model B – since both were regressed using the ISFF 2013's data. Remembering the results presented in the first section of this chapter, it is possible to conclude that both model A and B present, on average, better results than the other two models.

To conclude, Table 12 summarizes the accuracy rates of the models presented so far, with a cut-off equal to 15%, to give the notion of the different results:

	<i>Total accuracy</i>	<i>“Default accuracy”</i>	<i>“Non-default accuracy”</i>
<i>Model A</i>	74.02%	72.22%	74.28%
<i>Model B</i>	78.29%	66.00%	79.95%
<i>Henriques (2014)'s Model – Version 1</i>	61.34%	60.18%	61.54%
<i>Henriques (2014)'s Model – Version 2</i>	60.09%	69.35%	58.56%

Table 12: Accuracy rates of Model A, Model B, Henriques (2014)'s Model – Version 1 and Henriques (2014)'s Model – Version 2, with a cut-off equal to 15%

2. Model by Saunders and Cornett (2012)

In order to compare the mentioned probabilistic models with a different and simpler model, this section presents a comparison with the simple rating model developed by Saunders and Cornett (2012), in the United States. The main goal of the section is to investigate if a simple rating model can do the job.

The idea behind the model is more or less the same as a probabilistic or logistic regression, in which the model uses observed characteristics of the applicant to calculate a “score” that can be transformed into a PD (Saunders & Cornett, 2012). Using the borrowers personal and financial characteristics, the model weights each characteristic provided to identify a boundary number or range in which the applicant’s score must be higher than a predetermined score to be accepted for a loan. The theory behind the rating model is that, as stated by Saunders and Cornett (2012, p. 601), “*by selecting and combining different economic and financial characteristics, an FI manager may be able to separate good from bad loan costumers based on the characteristics of borrowers who have defaulted in the past*”.

The model proposed by these authors include the variables “annual gross income”; “total debt service”; “relations with the financial institution”, which translates the existence of a checking account, savings account or both; the existence of “major credit cards”; “age”; “residence”; “length of residence”; “job stability”; and “credit history”. As table 13 presents, according to each range and each variable, certain points are added. If the applicant’s total score is less than 120, the loan is automatically rejected; if the total score is higher than 190, the loan is automatically accepted; and if the total score ranges between 120 and 190, the loan is reviewed for a final decision.

<i>Characteristics</i>		<i>Characteristics' Values and Scores</i>				
<i>Annual gross income</i>		$\leq \$10,000$	$\$10,001 - \$25,000$	$\$25,001 - \$50,000$	$\$50,001 - \$100,000$	$> \$100,000$
	<i>Score</i>	0	15	35	50	75
<i>TDS</i>		$> 50\%$	$35\% - 50\%$	$15\% - 35\%$	$5\% - 15\%$	$< 5\%$
	<i>Score</i>	0	10	20	35	50
<i>Relations with FI</i>		None	Checking account	Savings account	Both	
	<i>Score</i>	0	30	30	60	
<i>Major credit cards</i>		None	1 or more			
	<i>Score</i>	0	20			
<i>Age</i>		< 25	$25 - 60$	> 60		
	<i>Score</i>	5	30	35		
<i>Residence</i>		Rent	Own with mortgage	Own outright		
	<i>Score</i>	5	20	50		

<i>Length of residence</i>	<i><1 year</i>	<i>1 – 5 years</i>	<i>>5 years</i>
<i>Score</i>	0	20	45
<i>Job stability</i>	<i><1 year</i>	<i>1 – 5 years</i>	<i>>5 years</i>
<i>Score</i>	0	25	50
<i>Credit history</i>	<i>No record</i>	<i>Missed a payment in the last 5 years</i>	<i>Met all payments</i>
<i>Score</i>	0	-15	50

Table 13: Variables, values and weights of the rating model developed by Saunders and Cornett (2012)

As it can be seen when analyzing the accuracy rate of the model¹¹ (Table 14), that was computed by applying the model automatically, without any kind of calibration (just a conversion from euros to U.S. dollars), the prediction accuracy for the default group of people is far from being a good one. The table shows that, of a total of 408 households that have defaulted in the past, the model only predicts 19 of them (translating into a “default accuracy” equal to 4.66%) and it asks for a second look on 58 of them.

<i>Saunders and Cornett (2012)’s Model</i>		<i>Observed</i>		
		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
<i>Estimated</i>	≤ 120	19	4	23
	121-189	58	13	71
	≥ 190	331	2,973	3,304
	<i>Total</i>	408	2,990	3,398
<i>Total accuracy</i>		88.05%		
<i>“Default Accuracy”</i>		4.66%		
<i>“Non-default accuracy”</i>		99.43%		

Table 14: Accuracy of Saunders and Cornett (2012)’s model with the conversion of the variable “total gross income” from EUR to USD, with a range between 120 and 190

These bad results were predicted, eventually because the model was developed for the United States in the year of 2012.

One good way to calibrate the model is to adjust the total gross income having in mind the purchasing power parity (PPP), instead of only converting euros to dollars. This is

¹¹ To compute the model, the variable “total gross income” was converted from USD to EUR, using the exchange rate on January 1st, 2013, since the data used was created in that year. At that time, 1.00000 EUR was equal to 1.32027 USD.

important because the cost of life is different from country to country and, therefore, the range of the variable “total gross income” could not be the most accurate one without this last adjustment. Having this in mind, the gross income of the households was adjusted, using information retrieved from OECD.Stat in which, in 2013, the national income per capita, in US dollars for the U.S. and Portugal, was \$53,933.632 and \$27,523.469, respectively (OECD Stat, 2013a, 2013b).

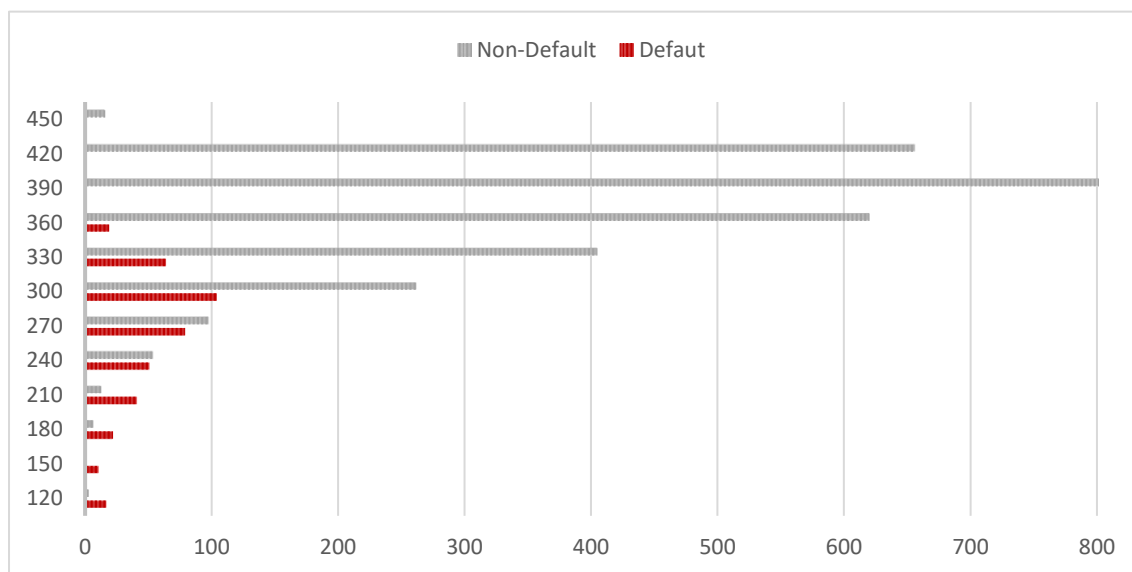
The following tables presents the accuracy of the Saunders and Cornett (2012)’s model, after the adjustment:

<i>Saunders and Cornett (2012)’s Model after the adjustment</i>		<i>Observed</i>		
		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
<i>Estimated</i>	≤ 120	17	3	20
	121-189	38	10	48
	≥ 190	353	2,977	3,330
	<i>Total</i>	408	2,990	3398
<i>Total accuracy</i>		88.11%		
<i>“Default Accuracy”</i>		4.17%		
<i>“Non-default accuracy”</i>		99.57%		

Table 15: Accuracy of Saunders and Cornett (2012)’s model with the adjustment of the variable “total gross income” using PPP with a range between 120 and 190

Despite the adjustment of the variable “total gross income”, having in consideration the income per capita on both countries, in PPP, in 2013, the results have barely changed. As it can be seen by analyzing Table 15, of a total of 408 defaults, the model is only able to predict 17 of them, representing an accuracy rate equal to 4.17%. This means that the model may be out of date, and a range calibration may be a good solution, since the remaining variables are more difficult to adjust.

The following graphs confirms this theory, where the vertical axis of the graph presents the scores and the horizontal axis presents the frequency of those scores considering the results of the adjusted model:



Graph 1: Frequency of the scores from the model of Saunders and Cornett (2012) after the adjustment of the variable "total gross income"

Points	120	150	180	210	240	270	300	330	360	390	420	450
Non-default	3	1	7	13	54	98	262	405	620	855	656	16
Default	17	11	22	41	51	79	104	64	19	0	0	0

Table 16: Frequency of the scores from the model of Saunders and Cornett (2012) after the adjustment of the variable "total gross income"

Both Graph 1 and Table 16 show that most households that have not defaulted in the past (represented in Graph 1 by the grey columns) present scores near 390 (855 households); while most of the default households (represented in Graph 1 by the blue columns) present scores near 300 (104 households), which is a difference of almost 100 points. Graph 1 also shows that less than 20 households that have defaulted present a score smaller than 120; while more than 3,000 households have scores higher than 190, which is the upper cut-off presented by the authors. This interpretation of the graph allows us to conclude that the model's range proposed by the authors in 2012 for the U.S. needs a calibration. Having this in mind, we propose different possible ranges, presented in Annex 11. It is important to note, though, that we decided to maintain the range between the minimum threshold and the maximum threshold proposed by the authors, which is equal to 70.

By calibrating the model until a range between 280 and 350 points, it is possible to achieve better results but not exactly ideal ones. With this interval, the model is able to predict

58.82% of the default clients and 62.71% of the non-default clients, representing a total accuracy rate equal to 62.24% (see Annex 11).

With the study of this model and its calibrations, it is possible to conclude that what seems to be a crucial step when developing a credit scoring/rating model is the definition of a good cut-off, or, in this case, a good range. A model may be very well defined but if a good cut-off is not well specified, the model loses its prediction accuracy.

At last, it is also possible to conclude that, even after a calibration of the Saunders and Cornett (2012)'s model, the credit scoring models presented in the last section – model A and B – presented higher accuracy rates.

Concluding and to further test the models proposed, the next chapter is focused on applying the model on other European countries' data, namely France, Spain and Italy, using the HFCS developed in 2013 in each country.

Chapter 5:

Application of the model on other European countries

Having access on data of 19 European countries (Austria, Belgium, Cyprus, Estonia, Finland, France, Greece, Hungary, Ireland, Italy, Latvia, Luxembourg, Malta, Netherlands, Poland, Portugal, Slovakia, Slovenia and Spain) – due to the HFCS database provided by the ECB -, this chapter intends to test the robustness of the model developed (model B). The expectations on the results are very high, since it will be possible to apply the model on 18 different countries, that will provide good insight on whether the model developed is robust, or not.

Unfortunately, after the analysis of the data of the countries, we concluded that it is not possible to apply the model on most countries due to missing observations on crucial variables, such as on the dependent variable, making completely impossible to reach any conclusions on whether the model has good accuracy results. Consequently, the only countries on which it is possible to apply the model is on France, Italy, Portugal and Spain, and only if some changes are made, namely removing from the model the variable “Having conditions deteriorated in the past 3 years”.

Having this in mind, a new model without this variable had to be created – model C – that, just like model B, follows a probabilistic regression, with the following expression:

$$\begin{aligned}
PD_i = Probability(default)_i &= \Phi (\beta_0 + \beta_1 Age_i + \beta_2 Level\ of\ Education_i + \beta_3 Marital\ Status_i \\
&+ \beta_4 Time\ at\ current\ job_i + \beta_5 Credit\ denied_i + \beta_6 Having\ savings_i \\
&+ \beta_7 Number\ of\ dependents_i \\
&+ \beta_8 Occupancy\ scheme\ 1: Total\ ownership_i \\
&+ \beta_9 Occupancy\ scheme\ 2: Co - ownership_i \\
&+ \beta_{10} Occupancy\ scheme\ 1: Rent_i + \beta_{11} Total\ Financial\ Assets_i \\
&+ \beta_{12} Expenses\ of\ the\ last\ 12\ months\ in\ relation\ to\ income\ 1: Superior_i \\
&+ \beta_{13} Expenses\ of\ the\ last\ 12\ months\ in\ relation\ to\ income\ 1: Inferior_i \\
&+ \beta_{14} Income_i + \beta_{15} Wealth\ (without\ financial\ assets)_i + \varepsilon_i)
\end{aligned}$$

where i is the family observed, β_0 is the constant of the model, β_1 up until β_{15} are the coefficients of the explanatory variables and ε is the error of the model that is not explained by the included variables.

Our first idea was to apply the model on each country data, individually, but that was not possible due to two reasons: first, it would not bring any significant results to Italy, due to the few number of observations tested (606 observations, from which 582 refer to non-defaulted clients and only 24 refer to the defaulted clients); and, second, it was not possible to regress the model using French data because the variable “Occupancy scheme: co-ownership” perfectly predicts binary response success, which means that this variable and the dependent one have practically the same outputs (they are both binary). These limitations lead me to regress the model using the data of the four countries together and to discriminate the country using dummies.

Just like in the other models regressed, the sample was first divided in two, one for training and one for testing. Therefore, the variables of the model are defined as follows:

<i>Model C</i>		
<i>Variable</i>	<i>Coefficient</i>	<i>Std. Error</i>
<i>C</i>	-0.494807 ***	0.175135
<i>Age</i>	-0.002213	0.002237
<i>Level of Education</i>	-0.196271 ***	0.057379
<i>Marital Status</i>	-0.105311 *	0.055603
<i>Time at current job</i>	-0.005014 **	0.002322
<i>Having savings</i>	-0.324035 ***	0.061503
<i>Number of dependents</i>	0.035383	0.027764

<i>Occupancy scheme 1: Total ownership</i>	-0.084020	0.123632
<i>Occupancy scheme 2: Co-ownership</i>	0.218290	0.186927
<i>Occupancy scheme 3: Rent</i>	-0.320341 **	0.136984
<i>Total financial assets</i>	7.35E-09	2.53E-08
<i>Expenses of the last 12 months in relation to income 1: Superior</i>	0.349002 ***	0.066616
<i>Expenses of the last 12 months in relation to income 2: Inferior</i>	-0.271018 ***	0.056179
<i>Income</i>	-7.93E-07 ***	2.58E-07
<i>Wealth (without financial assets)</i>	1.26E-08	1.23E-08
<i>Country: Portugal</i>	-0.065038	0.067440
<i>Country: France</i>	3.003019 ***	0.081939
<i>Country: Italy</i>	-0.652736 ***	0.107922
<i>McFadden R-squared</i>	0.584993	
<i>Akaike info criterion</i>	0.575672	
<i>Total observations</i>	5690	
<i>Observations with Dep=0</i>	3186	
<i>Observations with Dep=1</i>	2504	

Table 17: Model C

*, p-value < 0.1
 **, p-value < 0.05
 ***, p-value < 0.01

Comparing the results of this model with the others developed in the other chapters, table 18 presents the accuracy of the model in global terms, that is, the accuracy of the model

Model C		Observed		
Cut-off=15%		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 15\%$	2435	523	2958
	$< 15\%$	138	2647	2785
	<i>Total</i>	2573	3170	5743
Total accuracy		88.49%		
“Default Accuracy”		94.64%		
“Non-default accuracy”		83.50%		

Table 18: Accuracy of model C with a 15% accuracy rate, without discriminating the data of the countries

without discriminating between countries, with a cut-off equal to 15%¹².

Looking at Table 18, the results seem amazing. In fact, the accuracy of the default clients, which usually is not as high as 94%, is so great due to the nature of the French clients. If we look at Annex 13, it is possible to see that most of the clients from France seem to default (2161 from a total of 2281) which is unusual and not expected, at all. Therefore, the default accuracy is so high due to the French clients; and the non-default accuracy is high due to the clients of the other countries.

Analyzing Annex 13 that presents the results on an individual basis, it is possible to conclude that the countries that presented satisfactory results were Portugal and Spain. We can't conclude anything from the French and Italy database, due to the odd values on the first one; and the lack of observations on the second.

Then, in order to assess if it is better to regress the model using data from the countries together and discriminate the countries using dummies (and benefit from the higher number of observations); or to regress each country individually, Annex 14 shows the output of model C for Portugal and Spain, individually regressed, and the respective results in terms of accuracy. As it is possible to conclude, what seems to bring better results (higher accuracy rates) is to regress the model individually. As Annex 13 and 14 show, for a cut-off equal to 15%, the total accuracy of model C for Portugal is 78.23% when individually regressed (with a "default accuracy" equal to 62.00% and a "non-default accuracy" equal to 80.42%) versus 78.11% when regressed with the other countries (with a "default accuracy" equal to 55.00% and a "non-default accuracy" equal to 81.23%); and the total accuracy for Spain is 67.18% when individually regressed (with a "default accuracy" equal to 77.95% and a "non-default accuracy" equal to 65.03%) versus 59.32% when regressed with the other 3 countries together (with a "default accuracy" equal to 82.56% and a "non-default accuracy" equal to 54.69%).

With this chapter we were able to conclude that what seems to be the best is to regress the model individually to each country; and that it is better to benefit from the individual characteristics of each country when the model is regressed individually, than to benefit from a high number of observations.

¹² Annex 12 shows the accuracy of the model using different cut-offs (50%, 30%, 20% and 10%) and Annex 13 shows the results when discriminating the data of the countries, at different cut-offs (50%, 30%, 20%, 15% and 10%).

Chapter 6:

Conclusions

In this study, made in the context of a curricular internship done at Banco L. J. Carregosa, S.A., we developed a credit scoring model for the Portuguese private clients of the banking industry, based on a survey developed by the European Central Bank with conjunction with 20 European countries, in 2013, entitled “Household Finance and Consumption Survey”. The model, which follows a probabilistic regression, is able to estimate a probability of default, based on past data of 12 variables, namely: “age”, “level of education”, “marital status”, “time at current job”, “having savings”, “number of dependents”, “occupancy scheme”, “financial assets”, “expenses in relation to income”, “income”, “wealth” and “having conditions deteriorated in the past 3 years”. The choice of the variables to include in the model was made having in mind the most used ones in the literature, the ones present in the HFCS, and the ones that have passed the statistical tests conducted (significance test; mean difference and proportions test; and collinearity test). After its development, the objective was to evaluate it, in terms of accuracy, using five different cut-offs (50%, 30%, 20%, 15% and 10%), to conclude which one performs better; and then to compare it with other models developed in the literature, such as Henriques (2014)’s model and Saunders and Cornett (2012)’s rating model. At last, this study applied the model developed on European data, to test its robustness.

Despite the importance that the choice of the type of model to use has, as well as the variables to include in it, when developing a model, the use of five different cut-offs to evaluate the model enables the conclusion that the cut-off chosen plays an important role. As orally stated by Banco L.J. Carregosa, S.A.’s risk department members during the course of the internship made, a cut-off equal to 50% is very high for the purpose of identifying possible bad clients and does not bring satisfactory results. In fact, Banco Carregosa’s analysts tend to use one near to 10% or 15%. Their major focus is on the “default accuracy”,

since they prefer to avoid bad clients that won't pay instead of only focusing on the "non-default accuracy".

The model developed in this study – model B - presented good results, by being able to predict the behavior of 78.29% of the clients, with a "default accuracy" equal to 66.00% and a "non-default accuracy" equal to 79.95%. The comparison with other models in the literature, such as Henriques (2014) and Saunders and Cornett (2012), was important, since, with that study, we were able to achieve two conclusions: first, it is better to use models that are developed with more recent data; and, second, a credit scoring model, despite being more difficult to compute, brings better results when comparing to a simple rating model. Moreover, the application of the model on other European countries that, later, was only possible to do with Spain, allow us to test the robustness of the model; and to conclude that it is better to regress the model on each country individually, due to its better accuracy results. On the other hand, if the number of observations is not enough to regress a model (like what happened with Italy), the best is to regress together with other countries data.

These findings play an important role to financial institutes that may use this model on their daily work, since it was constructed using a database that is not public and is very complete, with important information regarding the household's personal and financial information, as well as their habits concerning investments and savings.

For future research, we recommend the application of this model on the next wave of the HFCS, especially the one developed by the Portuguese institutes involved (ISFF), provided by the Portuguese Statistic Institute, due to the limitations that the HFCS presented. Without these limitations – the lack of responses on crucial variables – we would be able to apply the model developed on more 16 countries, which would, without any doubt, provide some interesting results; as well as on private information that banks possess, to test it on their data. Another limitation that we faced during the development of the model was the anonymous character of the database. If the data was not anonymous, we would be able to recognize the households that have answered both waves of the survey – in 2010 and 2013 – allowing the development of a contemporary model, with a dependent variable at time t and independent variables at time $t-1$. This would bring value to the model, because, as stated by Avery et al. (2004, p. 524), "*a contemporary model* is built on the premise that past performance in repaying debts is the best prediction for future performance".

References

- Abdou, H., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: A review of the literature. *Intelligent Systems in Accounting, Finance and Management*, 18(2-3), 59-88.
- Alfaro, R., & Gallardo, N. (2012). The determinants of household debt default. *Revista de Análisis Económico*, 27(1), 55-70.
- Allen, L., DeLong, G., & Saunders, A. (2004). Issues in the credit risk modeling of retail markets. *Journal of Banking & Finance*, 28(4), 727-752.
- Altman, E., & Saunders, A. (1997). Credit risk measurement: Developments over the last 20 years. *Journal of Banking & Finance*, 21(11), 1721-1742.
- Anderson, R. (2007). The credit scoring toolkit: Theory and practice for retail credit risk management and decision automation. *OUP Catalogue*.
- Avery, R., Calem, P., & Canner, G. (2004). Consumer credit scoring: Do situational circumstances matter? *Journal of Banking & Finance*, 28(4), 835-856.
- Baker, H., & Filbeck, G. (2013). *Portfolio theory and management*: Oxford University Press.
- Banco de Portugal. (2017). Estatísticas dos empréstimos concedidos pelo setor financeiro. Retrieved from <https://www.bportugal.pt/page/estatisticas-dos-emprestimos-concedidos-pelo-setor-financeiro>
- Banco L. J. Carregosa. (2017). Informação a divulgar ao público sobre o incumprimento de contratos de crédito e a rede extrajudicial de apoio. Retrieved from <https://www.bancocarregosa.com/pt/repositorio/informacao-legal/risco-de-incumprimento.pdf>
- Bank for International Settlements. (2001). *The internal ratings-based approach*.
- Bank for International Settlements. (2006). *International convergence of capital measurement and capital standards: A revised framework*.
- Caouette, J., Altman, E., & Narayanan, P. (1998). *Managing credit risk: The next great financial challenge* (Vol. 2): John Wiley & Sons.
- Capon, N. (1982). Credit scoring systems: A critical analysis. *The Journal of Marketing*, 82-91.
- Constancioara, A. (2011). Consumer credit scoring. *Romanian Journal of Economic Forecasting*, 3, 162-177.
- Costa, S. (2012). *Probabilidade de incumprimento das famílias: Uma análise com base nos resultados do ISFF*. Departamento de Estudos Económicos do Banco de Portugal.
- Costa, S., & Farinha, L. (2012). Inquérito à situação financeira das famílias: Metodologia e principais resultados. *Banco de Portugal Occasional Papers*, 1, 2012.
- Crook, J., Edelman, D., & Thomas, L. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, 183(3), 1447-1465.
- European Securities and Markets Authority. (2016). *Competition and choice in the credit rating industry*. Retrieved from https://www.esma.europa.eu/sites/default/files/library/2016-1662_cra_market_share_calculation.pdf

- Everitt, B. (1992). *The analysis of contingency tables*. CRC Press.
- Fitch Ratings. (2017). *Rating Definitions*. Retrieved from <https://www.fitchratings.com/site/dam/jcr:6b03c4cd-611d-47ec-b8f1-183c01b51b08/Rating%20Definitions%20-%20March%2017%202017.pdf>
- Gonçalves, E., Gouvêa, M., & Mantovani, D. (2013). Análise de risco de crédito com o uso de regressão logística. *Revista Contemporânea de Contabilidade*, 10(20), 139-160.
- Greene, W. (1992). A statistical model for credit scoring.
- Hand, D., & Henley, W. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society Series a-Statistics in Society*, 160, 523-541.
- Henriques, C. (2014). *Modelo de notação de risco para famílias portuguesas*. (Master), Universidade Católica Portuguesa,
- Imtiaz, S., & Brimicombe, A. (2017). A better comparison summary of credit scoring classification. *Internatinal Journal of Advanced Computer Science and Applications*, 8(7), 1-4.
- Kocenda, E., & Vojtek, M. (2009). *Default predictors and credit scoring models for retail banking*. Working Paper Series No. 2862. CESifo Group Munich.
- Lima, J. (Producer). (2017, September 15). Portugal looks to attract new investors after S&P rating upgrade. *Bloomberg*. Retrieved from <https://www.bloomberg.com/news/articles/2017-09-15/portugal-looks-to-attract-new-investors-after-s-p-rating-upgrade>
- Mester, L. (1997). What's the point of credit scoring? *Business Review*, 3(Sep/Oct), 3-16.
- Moody's Investors Service. (2017). *Rating symbols and definitions*. Retrieved from https://www.moody's.com/researchdocumentcontentpage.aspx?docid=PBC_79004
- Obrova, V. (2012). *Construction and application of scoring models*. Karvina: Silesian Univ Opava, School Business Administration Karvina.
- OECD Stat. (2013a). Country statistical profiles: Portugal. Retrieved June 2nd 2018 <http://stats.oecd.org/index.aspx?queryid=58531>
- OECD Stat. (2013b). Country statistical profiles: United States. Retrieved June 2nd 2018 <http://stats.oecd.org/index.aspx?queryid=58539>
- Samreen, A., & Zaidi, F. (2012). Design and development of credit scoring model for the commercial banks of Pakistan: Forecasting creditworthiness of individual borrowers. *International Journal of Business and Social Science*, 3(17).
- Saunders, A., & Cornett, M. (2012). *Financial markets and institutions* (5th ed.): McGraw-Hill Irwin.
- Soares, R. (2017). Agressividade do crédito faz disparar riscos para as famílias. *Público*, 2017. Retrieved from <https://www.publico.pt/2017/12/10/economia/noticia/agressividade-do-credito-ao-consumo-faz-disparar-risco-para-as-familias-1795375>
- Sousa, M., Gama, J., & Brandão, E. (2016). A new dynamic modeling framework for credit risk assessment. *Expert Systems with Applications*, 45, 341-351. doi:10.1016/j.eswa.2015.09.055
- Šušteršič, M., Mramor, D., & Zupan, J. (2009). Consumer credit scoring models with limited data. *Expert Systems with Applications*, 36(3), 4736-4744.
- Thomas, L. (2000). A Survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 16(2), 149-172.
- Treacy, W., & Carey, M. (2000). Credit risk rating systems at large US banks. *Journal of Banking & Finance*, 24(1), 167-201.

- West, D. (2000). Neural network credit scoring models. *Computers & Operations Research*, 27(11), 1131-1152.
- White, L. (2002). The credit rating industry: An industrial organization analysis. In *Ratings, rating agencies and the global financial system* (pp. 41-63): Springer.

Annexes

Annex 1.: Market share calculation based on 2015 applicable turnover from credit rating activities and ancillary services in the EU (European Securities and Markets Authority, 2016).

<i>Registered Credit Rating Agency</i>	<i>Market Share</i>
<i>AM Best Europe-Rating Services Ltd. (AMBERS)</i>	0.93%
<i>ARC Ratings, S.A.</i>	0.03%
<i>ASSEKURATA Assekuranz Rating-Agentur GmbH</i>	0.21%
<i>Axesor S.A.</i>	0.05%
<i>BCRA-Credit Rating Agency AD</i>	0.02%
<i>Capital Intelligence (Cyprus) Ltd</i>	0.14%
<i>CERVED Group S.p.A</i>	0.88%
<i>Creditre Rating AG</i>	0.50%
<i>CRIF S.p.A</i>	0.05%
<i>Dagong Europe Credit Rating Srl</i>	0.04%
<i>DBRS Ratings Limited</i>	1.89%
<i>The Economist Intelligence Unit Ltd</i>	0.80%
<i>Euler Hermes Rating GmbH</i>	0.21%
<i>European Rating Agency, a.s.</i>	0.00%
<i>EuroRating Sp. Zo.o</i>	0.01%
<i>Feri EuroRating Services AG</i>	0.40%
<i>Fitch Group</i>	16.56%
<i>GBB-Rating Gesellschaft für Bonitätsbeurteilung mbH</i>	0.34%
<i>ICAP Group SA</i>	0.12%
<i>INC Rating Sp. Zo.o</i>	0.00%
<i>ModeFinance S.A.</i>	0.05%
<i>Moody's Group</i>	31.29%
<i>Rating-Agentur Expert RA GmbH</i>	0.00%
<i>Scope Ratings AG</i>	0.39%
<i>Spread Research SAS</i>	0.09%
<i>Standard & Poor's Group</i>	45.00%
<i>TOTAL</i>	100.00%

Annex 2.: Initial 68 variables considered.

<i>C of Credit</i>	<i>Definition</i>	<i>Variable</i>
<i>Personal and Socio-professional characteristics</i>	This category includes variables related to the reputation of the household, its willingness to repay as well as the personal characteristics of the household's representative.	<i>Personal characteristics & Educational background:</i> 1. Age; 2. Level of education; 3. Gender; 4. Marital status; 5. Level of education of the father; 6. Level of education of the mother; <i>Professional & Financial situation:</i> 7. Time at current job; 8. Credit application; 9. Credit denied; 10. Having a bank account; 11. Having a credit card; 12. Having a leasing contract; 13. Having savings; 14. Time at last job; 15. Situation at the current job; 16. Occupation; 17. Type of contract at current job; 18. Having another job; 19. Total time at any job; 20. Participation in a company; 21. Type of financial risk willing to assume; 22. Measure adopted to face having expenses higher than income, in the last 12 months; <i>Family situation:</i> 23. Number of dependents; 24. Number of people in the household; 25. Number of people in the household with a job; 26. Time at current address; 27. Home postcode; 28. Type of residence; 29. Residence: outer appearance; 30. Total residence surface; 31. Occupancy scheme; 32. Purchase mode of the current residence;

<i>Capital</i>	This category includes variables that may be seen as resources available to use when an undesirable and unpredicted situation happens.	33. <i>Current accounts;</i> 34. <i>Savings accounts;</i> 35. <i>Investment funds;</i> 36. <i>Treasury bonds;</i> 37. <i>Investments in a company;</i> 38. <i>Shares;</i> 39. <i>Accounts managed by clients' manager: other assets;</i> 40. <i>Value of credit conceded to friends and family;</i> 41. <i>Other financial assets;</i> 42. <i>Mutual funds;</i>
<i>Capacity</i>	This category includes variables related to the ability to repay as well as variables related to earnings volatility.	43. <i>Income;</i> 44. <i>Rent;</i> 45. <i>Time at maturity or time until the more recent renegotiation;</i> 46. <i>Monthly installment (including interest and amortizations);</i> 47. <i>Monthly installment of other loans;</i> 48. <i>Time at maturity or time until the more recent renegotiation of other loans;</i> 49. <i>Future income expectation;</i> 50. <i>Total expenses;</i> 51. <i>Expenses of the last 12 months in relation to income;</i> 52. <i>Expenses of the last 12 months in relation to the average;</i> 53. <i>Capacity to get financial support by friends and family;</i> 54. <i>Value of other expenses;</i> 55. <i>Expenses/Income;</i>
<i>Collateral</i>	This category includes assets that may be used as collateral.	56. <i>Debt;</i> 57. <i>Current residence value;</i> 58. <i>Current value of other residences;</i> 59. <i>Current value of what owns;</i> 60. <i>Current value of automobiles;</i> 61. <i>Current value of other vehicles;</i> 62. <i>Current value of high value objects;</i> 63. <i>Net value of participation in a company;</i> 64. <i>Wealth;</i>
<i>Cycle conditions</i>	This category includes variables related to the state of the business cycle.	65. <i>Had conditions deteriorated in the past 3 years;</i> 66. <i>Sector of the company where it has the main job;</i> 67. <i>Having conditions deteriorated in the next 2 years;</i> 68. <i>Year of acquisition of the main residence.</i>

Annex 3.: Variables which outliers were controlled, and respective minimum and maximums (before and after the winsorization process) and respective percentage of winsorization.

<i>N</i>	<i>Variable</i>	<i>Minimum €</i> <i>Lower Bound</i>	<i>Maximum €</i> <i>Upper Bound</i>	<i>% wins. €</i> <i>New</i> <i>Minimum</i>	<i>% wins. €</i> <i>New</i> <i>Maximum</i>
1	Time at current	0 years	55 years	1%	2%
	job	-22.5 years	37.5 years	0 years	36 years
2	Time at current	0 years	70 years	1%	2%
	address	-26.5 years	65.5 years	1 years	64 years
3	Number of	0 years	5 years	1%	2%
	dependents	-2 years	3 years	0 years	3 years
4	Total financial	€0.00	€2,740,182.00	1%	12%
	assets	-€41,610.25	€71,483.75	€0.00	€66,788.10
5	Income	€0.00	€3,802,500.00	1%	6%
		-€106,869.75	€308,944.25	€0.00	€300,000.00
6	Time at maturity				
	or time until the	3 years	55 years	2%	1%
	more recent	10 years	50 years	10 years	50 years
7	Time at maturity				
	or time until the	3 years	50 years	1%	1%
	more recent	5 years	45 years	5 years	40 years
8	Total expenses	€0.00	€34,950.00	1%	6%
		-€614.50	€2,997.00	€210.00	€2,900.00
9	Debt	€0.00	€1,084,000.00	1%	8%
		-€81,000.00	€135,000.00	€0.00	€130,192.32
10	Wealth	-€207,500.00	€20,747,892.00	1%	11%
		-€265,444.75	€541,249.25	-€19,573.25	€484,373.37

11	Expenses over	0.12%	23,157.89%	1%	6%
	income	-4.41%	35.77%	3.66%	35.07%

Annex 4.: Variables tested and respective results (the ones in red are the ones that were automatically excluded due to their results in any one of the tests or for not being available).

<i>Nº</i>	<i>Variable</i>	<i>Category</i>	<i>Sub-category</i>	<i>P-value</i> <i>(Test 1)</i>	<i>P-value</i> <i>(Test 2)</i>	<i>P-value</i> <i>(Test 3.1)</i>	<i>P-value</i> <i>(Test 3.2)</i>	<i>P-value</i> <i>(Test 4)</i>
1	<i>Age</i>	Socio-professional characteristics	Personal characteristics & Educational background	3.41%	0.54%	1.31%	***	****
2	<i>Level of education</i>	Socio-professional characteristics	Personal characteristics & Educational background	0.00%	**	**	**	0.00%
3	<i>Gender</i>	Socio-professional characteristics	Personal characteristics & Educational background	5.66%	**	**	**	3.29%
4	<i>Marital Status</i>	Socio-professional characteristics	Personal characteristics & Educational background	0.29%	**	**	**	0.21%
5	<i>Level of education of the father</i>	Socio-professional characteristics	Personal characteristics & Educational background	0.42%	**	**	**	*****
6	<i>Level of education of the mother</i>	Socio-professional characteristics	Personal characteristics & Educational background	0.61%	**	**	**	*****
7	<i>Higher level of education accomplished by the parents</i>	Socio-professional characteristics	Personal characteristics & Educational background	0.08%	**	**	**	*****
8	<i>Time at current job</i>	Socio-professional characteristics	Professional & Financial situation	0.01%	1.50%	0.00%	***	****

9	<i>Credit Application</i>	Socio-professional characteristics	Professional & Financial situation	27.47%	**	**	**	14.29%
10	<i>Credit denied</i>	Socio-professional characteristics	Professional & Financial situation	0.03%	**	**	**	0.16%
11	<i>Having a bank account</i>	Socio-professional characteristics	Professional & Financial situation	0.00%	**	**	**	0.17%
12	<i>Having a credit card</i>	Socio-professional characteristics	Professional & Financial situation	0.00%	**	**	**	0.00%
13	<i>Having a leasing contract</i>	Socio-professional characteristics	Professional & Financial situation	98.44%	**	**	**	48.22%
14	<i>Having savings</i>	Socio-professional characteristics	Professional & Financial situation	0.00	**	**	**	0.00%
15	<i>Time at last job</i>	Socio-professional characteristics	Professional & Financial situation	Variable not available				
16	<i>Situation at current job</i>	Socio-professional characteristics	Professional & Financial situation	0.00%	**	**	**	*****
17	<i>Occupation</i>	Socio-professional characteristics	Professional & Financial situation	Variable not available				

18	<i>Type of contract at current job</i>	Socio-professional characteristics	Professional & Financial situation	0.00%	**	**	**	0.26%
19	<i>Having another job</i>	Socio-professional characteristics	Professional & Financial situation	2.74%	**	**	**	0.31%
20	<i>Total time at any job</i>	Socio-professional characteristics	Professional & Financial situation	42.63%	2.80%	5.68%	***	****
21	<i>Participation in a company</i>	Socio-professional characteristics	Professional & Financial situation	94.60%	**	**	**	47.30%
22	<i>Type of financial risk willing to assume</i>	Socio-professional characteristics	Professional & Financial situation	6.26%	**	**	**	*****
23	<i>Measure adopted to face having expenses higher than income, in the last 12 months</i>	Socio-professional characteristics	Professional & Financial situation	Variable not available				
24	<i>Number of dependents</i>	Socio-professional characteristics	Family situation	0.00%	**	**	**	*****
25	<i>Number of people in the household</i>	Socio-professional characteristics	Family situation	0.00%	**	**	**	*****
26	<i>Number of people in the household with a job</i>	Socio-professional characteristics	Family situation	0.00%	**	**	**	*****

27	<i>Ratio between people with a job and people in the household</i>	Socio-professional characteristics	Family situation	0.00%	21.10%	***	0.00%	****
28	<i>Time at current address</i>	Socio-professional characteristics	Family situation	7.90%	0.09%	0.08%	***	****
29	Home postcode	Socio-professional characteristics	Family situation	Variable not available				
30	<i>Type of residence</i>	Socio-professional characteristics	Family situation	3.68%	**	**	**	*****
31	<i>Residence: outer appearance</i>	Socio-professional characteristics	Family situation	84.64%	**	**	**	*****
32	<i>Total residence surface</i>	Socio-professional characteristics	Family situation	0.00%	41.99%	***	0.00%	****
33	<i>Occupancy scheme</i>	Socio-professional characteristics	Family situation	0.00%	**	**	**	*****
34	<i>Purchase mode of the current residence</i>	Socio-professional characteristics	Family situation	21.86%	**	**	**	*****
35	<i>Financial Assets</i>	Capital	*	0.00%	0.00%	0.00%	***	****
36	<i>Income</i>	Capacity	*	0.00%	0.00%	0.00%	***	****
37	<i>Rent</i>	Capacity	*	Incorporated in another variable.				

38	<i>Time at maturity or time until the more recent renegotiation</i>	Capacity	*	25.06%	20.25%	***	41.17%	****
39	<i>Monthly installment (including interest and amortizations)</i>	Capacity	*	Incorporated in another variable.				
40	<i>Monthly installment of other loans</i>	Capacity	*	Incorporated in another variable.				
41	<i>Time at maturity or time until the more recent renegotiation of other loans</i>	Capacity	*	89.22%	28.87%	***	45.63%	****
42	<i>Future income expectation</i>	Capacity	*	44.81%	**	**	**	*****
43	<i>Total expenses</i>	Capacity	*	0.00%	0.10%	0.00%	***	****
44	<i>Expenses of the last 12 months in relation to income</i>	Capacity	*	0.00%	**	**	**	*****
45	<i>Expenses of the last 12 months in relation to the average</i>	Capacity	*	0.40%	**	**	**	*****
46	<i>Capacity to get financial support by friends and family</i>	Capacity	*	0.33%	**	**	**	0.25%
47	<i>Value of other expenses</i>	Capacity	*	Incorporated in another variable.				

48	<i>Expenses over income</i>	Capacity	*	0.00%	0.00%	0.00%	***	****
49	<i>Debt</i>	Collateral	*	65.73%	25.77%	***	9.43%	****
50	<i>Fixed Assets</i>	Collateral	*	0.00%	0.00%	0.00%	***	****
51	<i>Wealth</i>	Collateral	*	0.00%	0.59%	0.00%	***	****
52	<i>Wealth without financial assets</i>	Collateral	*	0.00%	0.37%	0.00%	***	****
53	<i>Had conditions deteriorated in the past 3 years</i>	Cycle Conditions	*	0.00%	**	**	**	0.00%
54	<i>Sector of the company where it has the main job</i>	Cycle Conditions	*	3.68%	**	**	**	*****
55	<i>Having conditions deteriorated in the next 2 years</i>	Cycle Conditions	*	1.66%	**	**	**	0.97%
56	<i>Year of the acquisition of the main residence</i>	5.74	*	1.66%	**	**	**	*****

*: The variable does not have a sub-category.

**: Test 2, Test 3.1 and Test 3.2 do not apply because the variable is a categorical one.

***: The test computed was either Test 3.1 or Test 3.2, depending the result of the Test 2.

****: Test 4 does not apply because the variable is continuous.

*****: Test 4 does not apply because the variable is categorical but non-binomial.

Annex 5.: Distribution of the variable “situation at current job”.

<i>Situation at the current job</i>		<i>Number of households</i>	<i>% of households</i>
(i)	<i>Regular paid worker</i>	3,206	51.64%
(ii)	<i>Worker on leave</i>	16	0.26%
(iii)	<i>Unemployed</i>	539	8.68%
(iv)	<i>Student</i>	12	0.19%
(v)	<i>Retired</i>	2,155	34.71%
(vi)	<i>Disabled</i>	71	1.14%
(vii)	<i>Domestic</i>	176	2.84%
(viii)	<i>Other inactive</i>	33	0.54%
	<i>Total</i>	6,208	100%

Annex 6.: Distribution of the variable “Sector of the company where it has main job”.

	<i>Sector of the company</i>	<i>Number of households</i>	<i>% of households</i>
(i)	<i>Agriculture, animal production, hunting, forest and fishing</i>	45	1.96%
(ii)	<i>Extractive and transformative industries, electricity, gas, steam, water, ..., waste management and decontamination</i>	419	18.27%
(iii)	<i>Construction</i>	166	7.24%
(iv)	<i>Wholesale, retail and vehicle repair</i>	291	12.69%
(v)	<i>Transportation and storage</i>	163	7.10%
(vi)	<i>Accommodation and catering</i>	117	5.10%
(vii)	<i>Communication and information services</i>	83	3.61%
(viii)	<i>Finance and insurance services</i>	109	4.75%
(ix)	<i>Public and defense administration</i>	370	16.13%
(x)	<i>Education</i>	237	10.33%
(xi)	<i>Health and social support</i>	205	8.94%
(xii)	<i>Artistic activities</i>	89	3.88%
	<i>Total</i>	2,294	100%

Annex 7.: Distribution of the variable “Year of the acquisition of the main residence”.

<i>Year of the acquisition of the main residence</i>		<i>Number of households</i>	<i>% of households</i>
(i)	1940-1950	27	0.53%
(ii)	1951-1960	87	1.74%
(iii)	1961-1970	222	4.44%
(iv)	1971-1980	565	11.29%
(v)	1981-1990	824	16.47%
(vi)	1991-2000	1,333	26.64%
(vii)	2001-2010	1,855	37.07%
(viii)	2011-2013	91	1.82%
<i>Total</i>		5,004	100%

Annex 8.: Accuracy of the models A and B, using cut-offs equal to 50%, 30%, 20% and 10%.

Model A		Observed		
Cut-off=50%		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 50\%$	2	1	3
	$< 50\%$	52	380	432
	<i>Total</i>	54	381	435
Total accuracy		87.82%		
“Default Accuracy”		3.70%		
“Non-default accuracy”		99.74%		

Accuracy of the model A with a 50% cut-off

Model A		Observed		
Cut-off=30%		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 30\%$	15	24	39
	$< 30\%$	39	357	396
	<i>Total</i>	54	381	435
Total accuracy		85.52%		
“Default Accuracy”		27.78%		
“Non-default accuracy”		93.70%		

Accuracy of the model A with a 30% cut-off

Model A		Observed		
Cut-off=20%		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 20\%$	29	61	90
	$< 20\%$	25	320	345
	<i>Total</i>	54	381	435
Total accuracy		80.23%		
“Default Accuracy”		53.70%		
“Non-default accuracy”		83.99%		

Accuracy of the model A with a 20% cut-off

Model A		Observed		
Cut-off=10%		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 10\%$	45	150	195
	$<10\%$	9	231	240
	<i>Total</i>	54	381	435
Total accuracy		63.45%		
“Default Accuracy”		83.33%		
“Non-default accuracy”		60.63%		

Accuracy of the model A with a 10% cut-off

Model B		Observed		
Cut-off=50%		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 50\%$	20	17	37
	$<50\%$	180	1464	1644
	<i>Total</i>	200	1481	1681
Total accuracy		88.28%		
“Default Accuracy”		10.00%		
“Non-default accuracy”		98.85%		

Accuracy of the model B with a 50% cut-off

Model B		Observed		
Cut-off=30%		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 30\%$	78	73	151
	$<30\%$	122	1408	1530
	<i>Total</i>	200	1481	1681
Total accuracy		88.40%		
“Default Accuracy”		39.00%		
“Non-default accuracy”		95.07%		

Accuracy of the model B with a 30% cut-off

Model B		Observed		
Cut-off=20%		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 20\%$	107	192	299
	$<20\%$	93	1289	1382
	<i>Total</i>	200	1481	1681
Total accuracy		83.05%		
“Default Accuracy”		53.50%		
“Non-default accuracy”		87.04%		

Accuracy of the model B with a 20% cut-off

Model B		Observed		
Cut-off=10%		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 10\%$	157	502	659
	$<10\%$	43	979	1022
	<i>Total</i>	200	1481	1681
Total accuracy		67.58%		
“Default Accuracy”		78.50%		
“Non-default accuracy”		66.10%		

Accuracy of the model B with a 10% cut-off

Annex 9.: Accuracy of Catarina Henriques (2014)’s model, using cut-offs equal to 50%, 30%, 20%, 15% and 10%.

Catarina Henriques (2014)’s Model Cut-off=50%		Observed		
		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 50\%$	17	38	55
	$< 50\%$	96	713	809
	<i>Total</i>	113	751	864
Total accuracy		84.49%		
“Default Accuracy”		15.04%		
“Non-default accuracy”		94.94%		

Accuracy of Catarina Henriques (2014)’s model with a 50% cut-off

Catarina Henriques (2014)’s Model Cut-off=30%		Observed		
		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 30\%$	34	112	146
	$< 30\%$	79	639	718
	<i>Total</i>	113	751	864
Total accuracy		77.89%		
“Default Accuracy”		30.09%		
“Non-default accuracy”		85.09%		

Accuracy of Catarina Henriques (2014)’s model with a 30% cut-off

Catarina Henriques (2014)’s Model Cut-off=20%		Observed		
		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 20\%$	52	200	252
	$< 20\%$	61	551	612
	<i>Total</i>	113	751	864
Total accuracy		69.79%		
“Default Accuracy”		46.02%		
“Non-default accuracy”		73.37%		

Accuracy of Catarina Henriques (2014)’s model with a 20% cut-off

Catarina Henriques (2014)'s Model Cut-off=15%		Observed		
		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 15\%$	68	289	357
	$< 15\%$	45	462	507
	<i>Total</i>	113	751	864
Total accuracy		61.34%		
“Default Accuracy”		60.18%		
“Non-default accuracy”		61.54%		

Accuracy of Catarina Henriques (2014)'s model with a 15% cut-off

Catarina Henriques (2014)'s Model Cut-off=10%		Observed		
		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 10\%$	86	424	510
	$< 10\%$	27	327	354
	<i>Total</i>	113	751	864
Total accuracy		47.80%		
“Default Accuracy”		76.11%		
“Non-default accuracy”		43.54%		

Accuracy of Catarina Henriques (2014)'s model with a 10% cut-off

Annex 10.: Accuracy of Catarina Henriques (2014)’s regressed model, using cut-offs equal to 50%, 30%, 20%, 15% and 10%.

Catarina Henriques (2014)’s Regressed Model Cut-off=50%		Observed		
		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 50\%$	0	0	0
	$< 50\%$	62	374	436
	<i>Total</i>	62	374	436
Total accuracy		85.78%		
“Default Accuracy”		0.00%		
“Non-default accuracy”		100.00%		

Accuracy of Catarina Henriques (2014)’s model with a new regression, with a 50% cut-off

Catarina Henriques (2014)’s Regressed Model Cut-off=30%		Observed		
		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 30\%$	12	22	34
	$< 30\%$	50	352	402
	<i>Total</i>	62	374	436
Total accuracy		83.49%		
“Default Accuracy”		19.35%		
“Non-default accuracy”		94.12%		

Accuracy of Catarina Henriques (2014)’s model with a new regression, with a 30% cut-off

Catarina Henriques (2014)’s Regressed Model Cut-off=20%		Observed		
		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 20\%$	33	92	125
	$< 20\%$	29	282	311
	<i>Total</i>	62	374	436
Total accuracy		72.25%		
“Default Accuracy”		53.23%		
“Non-default accuracy”		75.40%		

Accuracy of Catarina Henriques (2014)’s model with a new regression, with a 20% cut-off

Catarina Henriques (2014)'s Regressed Model Cut-off=15%		Observed		
		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 15\%$	43	155	198
	$< 15\%$	19	219	238
	<i>Total</i>	62	374	436
Total accuracy		60.09%		
“Default Accuracy”		69.35%		
“Non-default accuracy”		58.56%		

Accuracy of Catarina Henriques (2014)'s model with a new regression, with a 15% cut-off

Catarina Henriques (2014)'s Regressed Model Cut-off=10%		Observed		
		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 10\%$	48	216	264
	$< 10\%$	14	158	172
	<i>Total</i>	62	374	436
Total accuracy		47.25%		
“Default Accuracy”		77.42%		
“Non-default accuracy”		42.25%		

Accuracy of Catarina Henriques (2014)'s model with a new regression, with a 10% cut-off

Annex 11.: Accuracy rates of Saunders and Cornett (2012)’s model, using ranges between 240 and 310; 250 and 320; 260 and 330; 270 and 340; and 280 and 350.

<i>Saunders and Cornett (2012)’s Model</i>		<i>Observed</i>		
		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
<i>Estimated</i>	≤ 240	142	78	220
	241-309	209	417	626
	≥ 310	57	2495	2552
<i>Total</i>		408	2990	3398
<i>Total accuracy</i>		77.60%		
<i>“Default Accuracy”</i>		34.80%		
<i>“Non-default accuracy”</i>		83.44%		

Accuracy of Saunders and Cornett (2012)’s model with the conversion of the variable “total gross income” from USD to EUR, with a range between 240 and 310

<i>Saunders and Cornett (2012)’s Model</i>		<i>Observed</i>		
		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
<i>Estimated</i>	≤ 250	160	98	258
	251-319	202	497	699
	≥ 320	46	2395	2441
<i>Total</i>		408	2990	3398
<i>Total accuracy</i>		75.19%		
<i>“Default Accuracy”</i>		39.22%		
<i>“Non-default accuracy”</i>		80.10%		

Accuracy of Saunders and Cornett (2012)’s model with the conversion of the variable “total gross income” from USD to EUR, with a range between 250 and 320

Saunders and Cornett (2012)'s Model		Observed		
		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	≤ 260	195	137	332
	261-329	188	653	841
	≥ 330	25	2200	2225
	<i>Total</i>	408	2990	3398
Total accuracy		70.48%		
“Default Accuracy”		47.79%		
“Non-default accuracy”		73.58%		

Accuracy of Saunders and Cornett (2012)'s model with the conversion of the variable “total gross income” from USD to EUR, with a range between 260 and 330

Saunders and Cornett (2012)'s Model		Observed		
		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	≤ 270	221	176	397
	271-339	178	760	938
	≥ 340	9	2054	2063
	<i>Total</i>	408	2990	3398
Total accuracy		66.95%		
“Default Accuracy”		54.17%		
“Non-default accuracy”		68.70%		

Accuracy of Saunders and Cornett (2012)'s model with the conversion of the variable “total gross income” from USD to EUR, with a range between 270 and 340

Saunders and Cornett (2012)'s Model		Observed		
		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	≤ 280	240	263	503
	281-349	159	852	1011
	≥ 350	9	1875	1884
	<i>Total</i>	408	2990	3398
Total accuracy		62.24%		
“Default Accuracy”		58.82%		
“Non-default accuracy”		62.71%		

Accuracy of Saunders and Cornett (2012)'s model with the conversion of the variable “total gross income” from USD to EUR, with a range between 280 and 350

Annex 12.: Accuracy rates of model C with aggregated data from Portugal, France, Italy and Spain, for cut-offs equal to 50%, 30%, 20% and 10%.

<i>Model C with aggregated data</i> <i>Cut-off=50%</i>		<i>Observed</i>		
		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
<i>Estimated</i>	$\geq 50\%$	2158	122	2280
	$< 50\%$	415	3048	3463
	<i>Total</i>	2573	3170	5743
<i>Total accuracy</i>		90.65%		
<i>“Default Accuracy”</i>		83.87%		
<i>“Non-default accuracy”</i>		96.15%		

Accuracy rates of model C, with a cut-off equal to 50%

<i>Model C with aggregated data</i> <i>Cut-off=30%</i>		<i>Observed</i>		
		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
<i>Estimated</i>	$\geq 30\%$	2261	225	2486
	$< 30\%$	312	2945	3257
	<i>Total</i>	2573	3170	5743
<i>Total accuracy</i>		90.65%		
<i>“Default Accuracy”</i>		87.87%		
<i>“Non-default accuracy”</i>		92.90%		

Accuracy rates of model C, with a cut-off equal to 30%

Model C with aggregated data Cut-off=20%		Observed		
		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 20\%$	2375	523	2898
	$< 20\%$	198	2647	2845
	<i>Total</i>	2573	3170	5743
Total accuracy		87.45%		
“Default Accuracy”		92.30%		
“Non-default accuracy”		83.50%		

Accuracy rates of model C, with a cut-off equal to 20%

Model C with aggregated data Cut-off=10%		Observed		
		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 10\%$	2507	1486	3993
	$< 10\%$	66	1684	1750
	<i>Total</i>	2573	3170	5743
Total accuracy		72.98%		
“Default Accuracy”		97.43%		
“Non-default accuracy”		53.12%		

Accuracy rates of model C, with a cut-off equal to 10%

Annex 13.: Accuracy rates of model C discriminating the data from each country (Portugal, France, Italy and Spain), for cut-offs equal to 50%, 30%, 20% and 10%.

Model C: Portugal		Observed		
Cut-off=50%		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 50\%$	0	0	0
	$< 50\%$	200	1481	1681
	<i>Total</i>	200	1481	1681
Total accuracy		88.10%		
“Default Accuracy”		0.00%		
“Non-default accuracy”		100.00%		

Accuracy rates of model C, for Portugal, with a cut-off equal to 50%

Model C: Portugal		Observed		
Cut-off=30%		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 30\%$	36	28	64
	$< 30\%$	164	1453	1617
	<i>Total</i>	200	1481	1681
Total accuracy		88.58%		
“Default Accuracy”		18.00%		
“Non-default accuracy”		98.11%		

Accuracy rates of model C, for Portugal, with a cut-off equal to 30%

Model C: Portugal		Observed		
Cut-off=20%		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 20\%$	81	124	205
	$< 20\%$	119	1357	1476
	<i>Total</i>	200	1481	1681
Total accuracy		85.54%		
“Default Accuracy”		40.50%		
“Non-default accuracy”		91.63%		

Accuracy rates of model C, for Portugal, with a cut-off equal to 20%

Model C: Portugal		Observed		
Cut-off=15%		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 15\%$	110	278	388
	$< 15\%$	90	1203	1293
	<i>Total</i>	200	1481	1681
Total accuracy		78.11%		
“Default Accuracy”		55.00%		
“Non-default accuracy”		81.23%		

Accuracy rates of model C, for Portugal, with a cut-off equal to 15%

Model C: Portugal		Observed		
Cut-off=10%		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 10\%$	159	635	794
	$< 10\%$	41	846	887
	<i>Total</i>	200	1481	1681
Total accuracy		59.79%		
“Default Accuracy”		79.50%		
“Non-default accuracy”		57.12%		

Accuracy rates of model C, for Portugal, with a cut-off equal to 10%

Model C: France		Observed		
Cut-off=50%		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 50\%$	2158	120	2278
	$< 50\%$	3	0	3
	<i>Total</i>	2161	120	2281
Total accuracy		94.61%		
“Default Accuracy”		99.86%		
“Non-default accuracy”		0.00%		

Accuracy rates of model C, for France, with a cut-off equal to 50%

Model C: France		Observed		
Cut-off=30%		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 30\%$	2159	120	2279
	$< 30\%$	2	0	2
	<i>Total</i>	2161	120	2281
Total accuracy		94.65%		
“Default Accuracy”		99.91%		
“Non-default accuracy”		0.00%		

Accuracy rates of model C, for France, with a cut-off equal to 30%

Model C: France		Observed		
Cut-off=20%		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 20\%$	2159	120	2279
	$< 20\%$	2	0	2
	<i>Total</i>	2161	120	2281
Total accuracy		94.65%		
“Default Accuracy”		99.91%		
“Non-default accuracy”		0.00%		

Accuracy rates of model C, for France, with a cut-off equal to 20%

Model C: France		Observed		
Cut-off=15%		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 15\%$	2160	120	2280
	$< 15\%$	1	0	1
	<i>Total</i>	2161	120	2281
Total accuracy		94.70%		
“Default Accuracy”		99.95%		
“Non-default accuracy”		0.00%		

Accuracy rates of model C, for France, with a cut-off equal to 15%

Model C: France		Observed		
Cut-off=10%		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 10\%$	2160	120	2280
	$< 10\%$	1	0	1
	<i>Total</i>	2161	120	2281
Total accuracy		94.70%		
“Default Accuracy”		99.95%		
“Non-default accuracy”		0.00%		

Accuracy rates of model C, for France, with a cut-off equal to 10%

Model C: Italy		Observed		
Cut-off=50%		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 50\%$	0	0	0
	$< 50\%$	17	589	606
	<i>Total</i>	17	589	606
Total accuracy		97.19%		
“Default Accuracy”		0.00%		
“Non-default accuracy”		100.00%		

Accuracy rates of model C, for Italy, with a cut-off equal to 50%

Model C: Italy		Observed		
Cut-off=50%		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 30\%$	0	0	0
	$< 30\%$	17	589	606
	<i>Total</i>	17	589	606
Total accuracy		97.19%		
“Default Accuracy”		0.00%		
“Non-default accuracy”		100.00%		

Accuracy rates of model C, for Italy, with a cut-off equal to 30%

Model C: Italy		Observed		
Cut-off=20%		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 20\%$	0	0	0
	$< 20\%$	17	589	606
	<i>Total</i>	17	589	606
Total accuracy		97.10%		
“Default Accuracy”		0.00%		
“Non-default accuracy”		100.00%		

Accuracy rates of model C, for Italy, with a cut-off equal to 20%

Model C: Italy		Observed		
Cut-off=15%		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 15\%$	4	7	11
	$< 15\%$	13	582	595
	<i>Total</i>	17	589	606
Total accuracy		96.70%		
“Default Accuracy”		23.53%		
“Non-default accuracy”		98.81%		

Accuracy rates of model C, for Italy, with a cut-off equal to 15%

Model C: Italy		Observed		
Cut-off=10%		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 10\%$	4	36	40
	$< 10\%$	13	553	566
	<i>Total</i>	17	589	606
Total accuracy		91.91%		
“Default Accuracy”		23.53%		
“Non-default accuracy”		93.89%		

Accuracy rates of model C, for Italy, with a cut-off equal to 10%

Model C: Spain		Observed		
Cut-off=50%		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 50\%$	0	2	2
	$< 50\%$	195	978	1173
	<i>Total</i>	195	980	1175
Total accuracy		83.23%		
“Default Accuracy”		0.00%		
“Non-default accuracy”		99.80%		

Accuracy rates of model C, for Spain, with a cut-off equal to 50%

Model C: Spain		Observed		
Cut-off=30%		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 30\%$	66	77	143
	$< 30\%$	129	903	1032
	<i>Total</i>	195	980	1175
Total accuracy		82.47%		
“Default Accuracy”		33.85%		
“Non-default accuracy”		92.14%		

Accuracy rates of model C, for Spain, with a cut-off equal to 30%

Model C: Spain		Observed		
Cut-off=20%		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 20\%$	135	279	414
	$< 20\%$	60	701	761
	<i>Total</i>	195	980	1175
Total accuracy		71.15%		
“Default Accuracy”		69.23%		
“Non-default accuracy”		71.53%		

Accuracy rates of model C, for Spain, with a cut-off equal to 20%

Model C: Spain		Observed		
Cut-off=15%		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 15\%$	161	444	606
	$< 15\%$	34	536	570
	<i>Total</i>	195	980	1175
Total accuracy		59.32%		
“Default Accuracy”		82.56%		
“Non-default accuracy”		54.69%		

Accuracy rates of model C, for Spain, with a cut-off equal to 15%

Model C: Spain		Observed		
Cut-off=10%		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 10\%$	184	695	879
	$< 10\%$	11	285	296
	<i>Total</i>	195	980	1175
Total accuracy		39.91%		
“Default Accuracy”		94.36%		
“Non-default accuracy”		29.08%		

Accuracy rates of model C, for Spain, with a cut-off equal to 10%

Annex 14.: Output of model C when regressing individually for each country; and respective accuracy rates, for Portugal and Spain, for cut-offs equal to 50%, 30%, 20%, 15% and 10%.

<i>Variable</i>	<i>Model C: Portugal</i>		<i>Model C: Spain</i>	
	Coefficient	Std. Error	Coefficient	Std. Error
<i>C</i>	-0.478426 *	0.282676	0.593588 *	0.323725
<i>Age</i>	-0.003953	0.003899	-0.013898 ***	0.004377
<i>Level of Education</i>	-0.451910 ***	0.135988	-0.390243 ***	0.107221
<i>Marital Status</i>	-0.123028	0.097156	-0.388090 ***	0.108255
<i>Time at current job</i>	-0.003368	0.004335	-0.026774 ***	0.006475
<i>Having savings</i>	-0.508317 ***	0.095984	-0.276286 **	0.123189
<i>Number of dependents</i>	0.206677 ***	0.048590	0.035029	0.056704
<i>Occupancy scheme 1: Total ownership</i>	-0.219534	0.190365	-0.344089	0.233147
<i>Occupancy scheme 2: Co-ownership</i>	-0.037717	0.334051	-0.079610	0.313398
<i>Occupancy scheme 3: Rent</i>	-0.214161	0.220975	0.077776	0.268936
<i>Total financial assets</i>	1.16E-07	7.77E-07	2.11E-08	2.77E-08
<i>Expenses of the last 12 months in relation to income 1: Superior</i>	0.688657 ***	0.119348	0.631395 ***	0.106138
<i>Expenses of the last 12 months in relation to income 2: Inferior</i>	-0.076227	0.106346	-0.725189 ***	0.139696
<i>Income</i>	-1.89E-06 ***	6.83E-07	-4.61E-07	4.04E-07
<i>Wealth (without financial assets)</i>	1.17E-08	1.02E-07	2.16E-08	1.48E-08
<i>McFadden R-squared</i>	0.156186		0.225090	
<i>Akaike info criterion</i>	0.642042		0.754105	
<i>Total observations</i>	1703		1225	

Observations with <i>Dep=0</i>	1496	1005
Observations with <i>Dep=1</i>	207	220

Model C for Portugal and Spain, individually

**: p-value < 0.1*

****: p-value < 0.05*

****: p-value < 0.01*

Model C: Portugal individually Cut-off=50%		Observed		
		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 50\%$	16	8	24
	$< 50\%$	184	1473	1657
	<i>Total</i>	200	1481	1681
Total accuracy		88.58%		
“Default Accuracy”		8.00%		
“Non-default accuracy”		99.46%		

Accuracy rates of model C, for Portugal, individually regressed with a cut-off equal to 50%

Model C: Portugal individually Cut-off=30%		Observed		
		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 30\%$	69	71	140
	$< 30\%$	131	1410	1541
	<i>Total</i>	200	1481	1681
Total accuracy		87.98%		
“Default Accuracy”		34.50%		
“Non-default accuracy”		95.21%		

Accuracy rates of model C, for Portugal, individually regressed, with a cut-off equal to 30%

Model C: Portugal individually Cut-off=20%		Observed		
		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 20\%$	104	182	286
	$< 20\%$	96	1299	1395
	<i>Total</i>	200	1481	1681
Total accuracy		83.46%		
“Default Accuracy”		52.00%		
“Non-default accuracy”		87.71%		

Accuracy rates of model C, for Portugal, individually regressed, with a cut-off equal to 30%

Model C: Portugal individually Cut-off=15%		Observed		
		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 15\%$	124	290	414
	$< 15\%$	76	1191	1267
	<i>Total</i>	200	1481	1681
Total accuracy		78.23%		
“Default Accuracy”		62.00%		
“Non-default accuracy”		80.42%		

Accuracy rates of model C, for Portugal, individually regressed, with a cut-off equal to 15%

Model C: Portugal individually Cut-off=10%		Observed		
		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 10\%$	162	513	675
	$< 10\%$	38	968	1006
	<i>Total</i>	200	1481	1681
Total accuracy		67.22%		
“Default Accuracy”		81.00%		
“Non-default accuracy”		65.36%		

Accuracy rates of model C, for Portugal, individually regressed, with a cut-off equal to 10%

Model C: Spain individually Cut-off=50%		Observed		
		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 50\%$	34	33	67
	$< 50\%$	161	945	1106
	<i>Total</i>	195	978	1173
Total accuracy		83.46%		
“Default Accuracy”		17.44%		
“Non-default accuracy”		96.63%		

Accuracy rates of model C, for Spain, individually regressed, with a cut-off equal to 50%

Model C: Spain individually Cut-off=30%		Observed		
		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 30\%$	99	155	254
	$< 30\%$	96	823	919
	<i>Total</i>	195	978	1173
Total accuracy		78.60%		
“Default Accuracy”		50.77%		
“Non-default accuracy”		84.15%		

Accuracy rates of model C, for Spain, individually regressed, with a cut-off equal to 30%

Model C: Spain individually Cut-off=20%		Observed		
		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 20\%$	134	244	378
	$< 20\%$	61	734	795
	<i>Total</i>	195	978	1173
Total accuracy		74.00%		
“Default Accuracy”		68.72%		
“Non-default accuracy”		75.05%		

Accuracy rates of model C, for Spain, individually regressed, with a cut-off equal to 20%

Model C: Spain individually Cut-off=15%		Observed		
		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 15\%$	152	342	494
	$< 15\%$	43	636	679
	<i>Total</i>	195	978	1173
Total accuracy		67.18%		
“Default Accuracy”		77.95%		
“Non-default accuracy”		65.03%		

Accuracy rates of model C, for Spain, individually regressed, with a cut-off equal to 15%

Model C: Spain individually Cut-off=10%		Observed		
		<i>Default</i>	<i>Non-default</i>	<i>Total</i>
Estimated	$\geq 10\%$	167	447	614
	$< 10\%$	28	531	559
	<i>Total</i>	195	978	1173
Total accuracy		59.51%		
“Default Accuracy”		85.64%		
“Non-default accuracy”		54.29%		

Accuracy rates of model C, for Spain, individually regressed, with a cut-off equal to 10%