Links between Scores, Real Default and Pricing: Evidence from the Freddie Mac's Loan-Level Dataset

Maria Rocha Sousa, João Gama, Elísio Brandão

Abstract—**Evidence from the Freddie Mac's single loan-level dataset, first published in March 2013, shows that existing scores are effective to order individuals by risk, but they are not prepared to predict real default in each point in time.**

We investigate the dynamics and performance of over 16.7 million of fully amortized 30-year fixed-rate mortgages in the U.S, originated between 1999 and the first quarter of 2013. We identify the frailties of the frameworks used in default prediction, to draw implications to risk-based pricing designs. Analysis shows that not only scores diminished their ability to predict default when the mortgage crisis has come to public's attention, but also the real default rates by score are irregular over time. It is also apparent that, since 2009, lenders are firmly declining the subprime loans, and first year default rates have declined. There is a link between scores, lending and default, mostly influenced by the lending practices. There is a link between scores, default and pricing, but the mapping between them is far from being adequate.

Index Terms—**Freddie Mac, mortgage loan-level dataset, dynamics, credit risk, score, PD' misalignment, risk-based pricing.**

I. Introduction

The subprime mortgage lending crisis in the U.S. came to public's attention when home foreclosures begun to rise in 2006 and moved out of control in 2007. A large decline in home prices prompted a devaluation housing-related securities and an unprecedented rise in mortgage delinquencies. This brought into light the disproportionate risk assumed in mortgage lending in the last decade, along the bursting of the U.S. housing bubble, between 2001 and 2005. This crisis echoed severely in the financial arena and in real economies worldwide. The collapse of several major financial institutions in 2008, promoted the distrust inside the financial systems. As a consequence, banks' liquidity plummeted with a significant disruption of the financing of businesses and consumers. Thus far, the U.S. and European Communities are still recovering from a severe recession. This spawned intensive debates towards causes and possible remedies, in view of achieving transparency and global financial stability.

Since 21 March 2013, Freddie Mac is making available loan-level credit performance data on a portion of fully amortized 30-year fixed-rate mortgages that the company purchased or guaranteed since 1999. This had never been done before by a loan level agency. The data is provided in a "living" dataset [1]. By June 2014, the dataset covers over 16.7 million of fully amortized, 30-year fixed-rate mortgages in the U.S., originated between 1999 and the first quarter of 2013. These loans represent a total amount granted of over 3,020 US B$. Disseminating these data follows the direction of the regulator, the Federal Housing Finance Agency (FHFA), as a part of a larger effort to increase transparency and promote risk sharing. The primary goal of turning this data available is to help investors build more accurate credit performance models in support of the risk sharing initiatives highlighted by the FHFA in the 2013 conservatorship scorecard [2]. The availability of such a large real world financial dataset also creates an unprecedented opportunity for researchers and practitioners, as it substantiates a more profound investigation on the roots of the global crisis. The aggregated data summary statistics is updated by Freddie Mac [3].

Anderson, Scott, and Janet Jozwik [4] propose a framework for developing a credit model based on this dataset. For a 180-days delinquent target event, the authors conclude that much of the variation in credit performance across loans and over different stages of the economic cycle is explained by loan-level variables. Unsurprisingly, by adding factors to capture broader macroeconomic effects and the quality of underwriting, they significantly improve the model. Goodman, Landy, Ashworth and Yang present an exploratory paper [5] providing a first look through the data, to find potential implications for guarantee pricing. The authors show the vintage composition as a percentage of the initial balance in a cross-analysis of the original borrowers' FICO score by the original loan to value (LTV). They follow the cumulative default in three groups in the score ranges 300 to 700, 700 to 750 and 750 to 850 crossed by the original LTV in selected buckets. They conclude that default rates are dramatically higher on higher LTV/lower scores, and so, investors should look not only at the average LTV and FICO scores, but also at the FICO/LTV loans' distribution. The authors conjecture that pricing these pools by looking at averages are likely to lead to underpriced default risk, but they do not present evidence.

Discussion is being pushed towards risk-based pricing. Previous studies suggest that risk-based pricing models will rely mostly in credit scores. This research extends the existing published work by proving meaningful insights on the link between credit score, lending practices, real default and pricing. Our research addresses the question: is there is a link between scores, real default and pricing?

Our research confirms that there is a link between scores, default and pricing, but the mapping between them is far from being adequate. New evidence from the Freddie Mac's single

Manuscript received September 5, 2014; revised December 13, 2014.

The authors are with the School of Economics and Management, University of Porto, Portugal (e-mail: 100427011@fep.up.pt, jgama@fep.up.pt, ebrandao@fep.up.pt).

DOI: 10.7763/JOEBM.2015.V3.343

loan level data shows that although existing scores consistently rank order portfolios' risk, real default rates by score are irregular over time. This means that existing scores are effective to rank order risk, but they are not prepared to adapt predictions to real default in each point in time.

This paper follows in section II with a formalization of the problem and a description of the research background. In section III, we present an overview of credit scoring models. First, we review the current role of credit scoring in the advanced economies, and then we present credit scoring formulation, for an in-depth comprehension. The section ends with a brief explanation of current capabilities and potential frailties of credit scoring models when they are used at the basis of credit risk underwriting and risk-based pricing. Experimental design is explained in section IV, and results are provided in section V. Selected outcomes are presented in order to illustrate dynamics over time in focusing the dimensions in analysis - score buckets, lending practices, default and pricing. Conclusions are drawn in sectionVI.

II. Problem and Research Background

A. The Problem

Lenders determine if the risk of lending to a borrower is acceptable under certain parameters of credit risk, borrower's credit capacity and collateral evaluation. Nowadays, in retail lending, a great proportion of the loan applications are automatically evaluated. In this setting, credit score is the central, if not unique, indicator of the borrowers' credit risk, either when the credit decision assessment is fully automatic or when it is an input for human decision. An individual without a credit score or with a low score (meaning high risk) is unlikely to have credit, whilst an application of a person with a high score has good chances to be accepted. An analysis on the causes and effects of the mortgage meltdown [6] states that in 2007, 40% of all subprime loans have been generated by automatic underwritings in the U.S. This had been associated to lax controls in the underwriting processes. Automated processes meant fasters decision, but less documentation scrutiny. Another important conclusion is that the acceptance standards rapidly moved towards credit score's over-dependence. Hence, the performance of credit loans is mostly reliant on the credit scoring models accuracy, both in the short-term as in the long-run predictions, which is too hard to achieve. In 2007, the delinquency rate rose sharply, both in borrowers in the lower scores as in the highest scores bands, showing that the actual risk of these borrowers have been underestimated.

Enhancing loans risk-based pricing models in the track of the previous studies [4], [5] will much depend on the knowledge and ability to improve the existing credit scoring robustness. This entails a deeper understanding of their actual strengths and current frailties.

B. Research Background

Research in credit risk assessment often lacks from validation in representative real world environments, and most of the experimental designs use datasets that are not representative of each phase in the economic cycles. Hence, a significant portion of empirical studies have no generalization ability. In particular, trying to screen the credit losses and predicting credit default of future credit operations may become critical if there is neither sufficient knowledge of the past neither of the future of the potential circumstances. In this setting, theoretical contributions have a more limited space to influence real world decisions and subjective and uninformed reasoning may prosper.

The unavailability of representative loan-level datasets has shortened the space to turn evident in which conditions the existing credit scoring models may be ineffective, like biased credit policies, drifting population and recessions. The single-family mortgage loan level dataset creates an unprecedented opportunity for researchers and practitioners, as long as it substantiates the simulation of theoretical frameworks in a real-world stressed environment.

III. Credit Scoring Fundamentals and Current Use

A. Credit Scores - A Standard Risk Assessment Measure in the Advanced Economies

Financial industry turned over-dependent of credit scoring over the last few decades. The origin of these models traces back the World War II, which promoted the first expert systems to evaluate a person's credit worthiness. As credit analysts were called to fight, finance houses and mail-order firms requested them to write down their rules for deciding whom to give loans. Some of these were numerical scoring systems and others were sets of conditions that needed to be satisfied – expert systems. In the early 1950s, Bill Fair and Earl Isaac created the first consultancy directed to finance houses, retailers and mail-orders firms, making use of statistically derived models in lending decision. Until 1970 credit risk assessment relied most exclusively in human judgment. Connected with the lending activities, this task was typically performed to support decision-making, following a specific credit application. The labor of the person responsible for the evaluation, often a branch manager, would involve the analysis of the likelihood of a customer repaying his debt, based on a number of clues that the manager could gather on site from a community leader or an external entity, such as the employer, a credit bureau or even another lender. Main aspects that he would check would concern to the customer character and honesty and his ability to create wealth. The depth of reasoning behind a decision could largely vary and the final decision would likely depend on the evaluator's mood and instinct. From customer application to the decision or credit granting, the process was usually slow [7]. Nowadays, scoring models are used in credit approval, risk management, internal capital allocation and in corporate governance functions of banks using the IRB approach. In the U.S., since its introduction 20 years ago, FICO score is calculated from the information available in the individuals' credit bureau reports, and has become an industry standard. It is claimed to be used in 90% of lending decisions, to determine how much money each individual can borrow, and how much interest he will pay. In the OECD countries, banks that have adopted the Internal Rating Based Approach (IRB) in Basel II, internally developed credit scoring models play an essential role in the calculation of the minimum regulatory capital. In the European Market, there are 89 banks using the IRB. In the US, the largest banks also adopted the Basel II Accord, introduced via Capital Requirements Directive. In line with this evolution, financial industry moves toward a more intensive use of credit scores at the basis of risk-based pricing models.

B. Credit Score Formulation

A credit scoring model is a simplification of the reality. The output is a prediction of a given entity, actual or potential borrower, entering in default in a given future period. Having decided on the default concept, conventionally a borrower being in arrears for more than 90 days in the following 12 months, those matching the criteria are considered bad and the others are good. Other approaches may consider a third status, the indeterminate, between the good and the bad classes, e.g. 15 to 90 days overdue, for which it may be unclear whether the borrower should be assigned to one class or to the other. This status is usually removed from the modeling sample; still the model can be used to score them.

The output of these models is a function of the input characteristics $\mathbf{x}$, which is most commonly referred as score, s($\mathbf{x}$). We also consider that this function has a monotonic decreasing relationship with the probability of entering in default (i.e. reaching the bad status). A robust scorecard enables an appropriate differentiation between the good and the bad classes. It is achieved by capturing an adequate set of information for predicting the probability of the default concept (i.e. belonging to the bad class), based on previous known default occurrences. The notation of such probability, Pr{bad|score based on X}, is:

$$p(B|s(\mathbf{x})) = p(B|s(\mathbf{x}), \mathbf{x}) = p(B|\mathbf{x}), \forall \mathbf{x} \in \mathbf{X} \quad (1)$$

Since p(G|$\mathbf{x}$)+p(B|$\mathbf{x}$)=1, it naturally follows the probability of the complementary class:

$$p(G|s(\mathbf{x})) = P(G|\mathbf{x}) = 1 - p(B|\mathbf{x}), \forall \mathbf{x} \in \mathbf{X} \quad (2)$$

Among researchers and real-world applications, a usual written form of the score is the log odds score:

$$s(\mathbf{x}) = \ln \frac{p(G|\mathbf{x})}{p(B|\mathbf{x})} \text{ and } p(B|\mathbf{x}) + p(G|\mathbf{x}) = \mathbf{1} \quad (3)$$

In so saying, the score may vary from -∞, when $P(G|\mathbf{x}) = 0$, to +∞, when $P(G|\mathbf{x}) = 1$, i.e. s($\mathbf{x}$) ∈ R.

The probability of the default event can be written in terms of the score:

$$p(B|\mathbf{x}) = \frac{1}{1+e^{s(\mathbf{x})}}, \forall \mathbf{x} \in \mathbf{X} \quad (4)$$

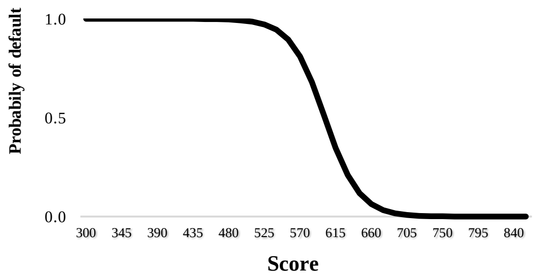
A conventional way to produce log odds score is based in the logistic regression. However, other classification algorithms can also be used, adjusting the output to the scale of that function. Then, we may assume that, independently of the method used to determine the best separation between the two classes, good and bad, and the resulting scorecard has the same property of the log odds score. Although a grounded mathematical treatment may be tempting to tackle this problem, it goes beyond the scope of this work. The basics of credit scoring and the most common approaches to build a scorecard, are further detailed in the operational research literature [7], [8]. Recent advances in the area deliver methods to build risk-based pricing models [9] and methodologies towards the optimization of the profitability to the lenders [10].

C. Scoring Models – Strengths and Frailties Strengths

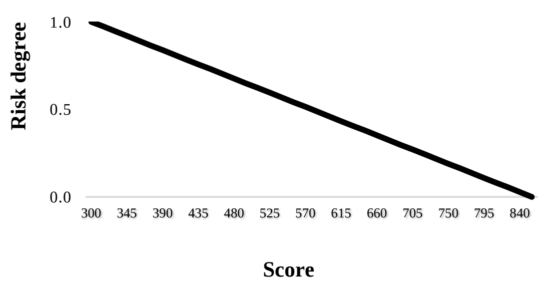
The enormous success of credit scoring models in the advanced economies is partly explained by their appealing representation is a linear scale. Credit scoring models have also proved to a powerful measure to rank order a population of individuals according to their credit risk. The best scoring model is the one that differentiates the most the two target classes, good and bad, commonly referred as discriminatory power.

D. Human Misconception

Although the true meaning of scores is a probability of default (PD) with a non-linear shape, as in Fig.1(a), human cognition retains the linear representation, Fig.1(b), rather than the actual non-linear shape. Through our experience in developing credit scoring models it became apparent that many of the risk managers were basing their credit risk assessments on the linear representation. Doing so is suitable for ranking risks, but it is insufficient to calculate losses or pricing credit risks.



(a) Actual meaning - Probability of default by score.



(b) Misconception - Human cognition of risk by score.

Fig. 1. Actual meaning of scores (top figure) and human misconception of the risk (bottom figure). Authors' analysis, illustrative scorecard: the score 660 corresponds to a good/bad odd of 15 good individuals to 1 bad, and for each additional 15 score points the good/bad odds double.

E. Economic Cycle and Scores' Pd Misalignment[2]

Traditional systems that are the basis of credit scoring

[2] The terms alignment, adjustment and calibration are commonly used with the same meaning.

models are one-shot, fixed memory-based, trained from fixed training sets. Since static models are not prepared to process the highly detailed evolving data. They are not able to continuously maintain an output (PD for scores) consistent with the actual state of nature, or to quickly react to changes [11]. When there are significant changes in the conditions, scores' PD (Fig. 2, dashed line) may become misaligned with the real default (Fig. 2, solid line).

As the processes underlying credit risk are not strictly stationary, consumers' behavior and default can change over time in unpredictable ways. There are several types of evolution inside a population, like population drifts, that translate into changes in the distributions of the variables, affecting the performance of the models.

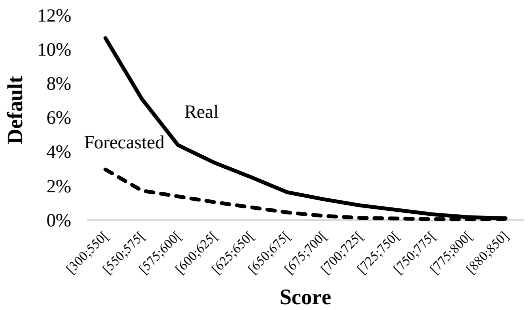


Fig. 2. Example of scores' PD misalignment. Authors' analysis. Forecasted PD's by score same as the default rate of the mortgages originated in 1999, and observed PD's the real default rates in 2007. Performance measured in the 1-year after the loans were originated.

As the economic conditions evolve in the economic cycle, either deteriorating or improving, also varies the behavior of an individual, and his ability of repaying his debt. In addition, default evolution echoes trends of the business cycle, and related with this, regulatory movements, and interest rates fluctuations. In good times, banks and borrowers tend to be overoptimistic about the future, whilst in times of recession banks are swamped with defaulted loans, high provisions, and tighten capital buffers turn highly conservative. The former leads to more liberal credit policies and lower credit standards, the later promotes sudden credit-cuts. Empirical evidence and theoretical frameworks support a positive, and lagged relationship between rapid credit growth and loan losses.

In order to adapt models' output to changes over time, institutions should calibrate their scoring models according to the most recent information. Models' adjustments, or calibration, commonly consider selected macroeconomic public indicators and should be periodically revised. In so doing, resulting adjusted scores translate a combination of the customers' specific risk with systemic risk. However, this may take too long to occur. The European Banking Authority reports that there is not a common practice among Regulators towards models calibration. Many countries do not define any specific rules and, when they do, they are usually not public. When they define some rules, they are rarely convergent; and different countries favor different calibration choices [12].

Should there be significant changes in between scheduled modeling developments or adjustments, it is not certain that banks will anticipate any of these tasks, as it can largely depend on judgmental reasoning or over-layered decision frameworks. Given that the time to decide or adapt may take too long to occur, the aim of opportunely adjusting rating systems, or credit policies, may become disappointing.

In the following, we will analyze the Freddie Mac's database to find evidence on the previous limitations. Complementing the previous works, we provide a fine-graded time-analysis over scores and analyze in which conditions risk-based pricing has been implemented. Furthermore, we study the extent of misalignment of PD's with the real default by score over time.

IV. EXPERIMENTAL DESIGN

The research summarized here was conducted in the Freddie Mac's single family mortgage loan-level dataset, first published in March 2013. We follow the performance of 16.737 million of fully amortized 30-year fixed-rate mortgages loans in the U.S., originated between January 1, 1999 and March 31, 2013. The loans performance [3] is outlined in a monthly basis and, at the time of this research, data for performing loans and those that were up to 180 days delinquent were available through September 30, 2013.

The dataset is a "living" dataset updated over time, typically at the end of each quarter, and may be subjected to periodical corrections by Freddie Mac. The release changes are recorded [13]. A general user guide describing the file layout and data dictionary is also provided [14].

Freddie Mac's information regarding the key loan attributes and performance metrics can be linked to our research in the aggregated summary statistics.

A. Methodology

We attempt to describe the most significant events in the period. We are both interested in determining the main contrasts by score, and to illustrate the dynamics over time. First, we illustrate the volumes and compare the original interest rate with the annual average FIX 30. In so doing, our aim is to use a representation of the credit risk spread evaluation over time and understand the extent of underpriced loans that has been referred as one cause of the crisis. Then, we determine the evolution of default over time and the performance of the scores at the basis of the credit risk assessment. For assigning the default event, we used the information of the loan delinquency status in each reporting period. In this analysis we consider that a borrower entered in default if he was ever 90 or more days delinquent, the typical definition used under the Basel II. Default is assigned to the first occurrence of this event. We use vintage analysis and hence, we consider cumulative default along time.

[3] Loan performance information includes the monthly loan balance, delinquency status and information regarding termination events: Voluntary prepayments in full; 180 days delinquency ("D180"); Repurchases prior to D180; Third-party sales prior to D180; Short sales prior to D180; Deeds-in-lieu of foreclosure prior to D180; Real estate owned (REO) acquisition prior to D180.

Specific credit performance information in the dataset includes voluntary prepayments and loans that were short sales, deeds-in-lieu of foreclosure, third party sales, and REOs.

B. Data Aggregation

Although we have found missing values in the score, we intentionally kept these cases in the analysis to entirely represent the extent of information (or absence) at the basis of the original risk assessment. Data of the original datasets were aggregated by the origination year.

Scores may vary in the range 301-850, or be unknown. Situations where the score is unknown are described by Freddie Mac [14][4]. To compare evolutions over time, we divided the range into equidistant intervals of 25, except for the lower and upper bounds. To have dimension, these bounds we aggregated in the buckets [300, 550] and [800, 850], respectively.

C. Scores - Performance, Concentration and Stability Measures

The discriminatory power of the model was measured based on the Gini coefficient, equivalent to consider the area under the ROC curve (AUC), which is a typical evaluation criteria among researchers and in the industry [15]. This coefficient refers to the global quality of the credit scoring model, and may range between -1 and 1. The perfect scoring model fully distinguishes the two target classes, good and bad, and has a Gini index equal to 1. In fact, this situation does not occur in practice, because it would represent the certain event. A model with a random output has a Gini coefficient equal to zero. If the coefficient is negative, then the scores have a reverse meaning. The extreme case -1 would mean that all examples of the good class are being predicted as bad, and vice-versa. In this case, the perfect model can be achieved just by switching the prediction. Loans' records with unknown score were not included in the calculations of this indicator.

The concentration of loans by score buckets was measured with the Herfindahl-Hirschman Index (HHI), which is defined as $\sum_{i=1}^{n} f_i^{\,2}$, where n is the number of score buckets and f_i is the number of customers in that bucket relative to the total portfolio for which the scores are known. By definition, the index varies between 0 and 100%. An index of 100% means that customers are concentrated in a single bucket. In this work we will consider that values below 20% are commonly acceptable. Values above it suggest a highly concentrated scores.

The stability index was measured comparing the distribution of the population in each year with the distribution of the population in the first year in the period, 1999. For the year Y, $Y = 2000, .., 2013$ it is calculated as:

$$\sum_{i=1}^{n} \left(f_{i,1999} - f_{i,Y}\right) * ln\left(f_{i,1999}/f_{i,Y}\right)$$

where n is the number of score buckets, and $f_{i,y}$ is the number of customers in that bucket relative to the total portfolio for which the scores are known in the year y.

[4] A possible reason is when the seller requires a reduced level of verification.

V. Results

There are a number of theories regarding the origins of the mortgage crisis. Our research is concerned to extend knowledge on the potential misalignment of the risk indicators that were at the basis of credit approval before crisis. Then, assess the extent of undervaluation of the credit risk, and hence, credit risk mispricing. As credit risk assessment is anchored in the borrowers' score at the origin of the loan, the analysis is focused in the dimensions score and time. Results are mostly motivated to present evidence on the following:

- Jeopardizing practices regarding risk taking and mispricing;
- Changes in the lending practices after the crisis;
- Risk assessment over time - default rates, default misalignment and scores performance.

A. Jeopardizing Practices in Risk Taking and Mispricing

The evolution of new loans over the analyzed period illustrates the U.S. housing bubble between 2001 and 2005. Higher peaks occur between 2001 and 2003, where the numbers of new loans continuously rose from nearly 800 thousand new loans in 2000 to 1,930 thousands in 2003 (Table I, 1st row). This massive increase in the volumes was one of the sources of the raise in the real state property values that reached a peak by 2005.

In the entire period, the analysis confirms that the scores are used to differentiate the interest rates of the mortgages. As Table II illustrates, there is a decreasing trend of the average interest rate from the lower to the higher score buckets. It can be said that a risk-based pricing based in scores is being applied. Until 2009, borrowers with the highest scores were borrowing below the average FIX 30. By the crisis, the default of the borrowers in the highest scores' borrowers has tripled in relation to the previous years. This suggests that credit risk is these borrowers may has been underpriced. Adjustments to the risk premium were made by 2009. From this point onwards, rates are higher than FIX 30, suggesting that pricing policy had been revised or the default forecast had been adjusted to higher values. Loans have been priced below the FIX 30 in 2001, both in the aggregate as in each score bucket (Table I, rows 7, 8 and 9). Loans' average rate was maintained through 2001 and 2002, in the aggregate level (Table II) and in score buckets' level (Table II). This effect may be linked to the crash of the dot-com bubble in 2000 which has been associated to the beginning of the decline in real long-term interest rates [6]. In reaction to the crash of the dot-com bubble in 2000 and to the recession that began in 2001, the Federal Reserve Board cut short-term interest rates from 6.5% to 1%. The mortgage interest rates continued to decline until 2005 (Table II). As the mortgage rates are typically set relation to 10-year Treasury bond yields, this was an outcome of very low Fed funds' rates in the period. Lenders were self-reliant that they were taking little risk because the value of the collateral was rising too fast, but they missed to understand that it would come to an end. Loans underwritten between 2001 and 2005 account for 42% of the amount originated in the period (Table IV). There is a theory [16] referring that in this period, lenders had begun to take more risk in subprime 5 mortgages. Our analysis provides a divergent finding, because neither the number of loans or amount granted has increased during the mortgage bubble (Table III and Table IV, see rows for scores' buckets bellow score 625).

B. Changes in the Lending Practices after the Crisis

We provide evidence on that major drifts are occurring in the lending practices as a reaction to the crisis, both in the acceptance scores thresholds as in the underlying credit risk spreads (Fig. 3). From 2009 onwards, rates have increased in all scores, when compared to the average FIX 30 in the year.

Borrowers' score is a key indicator in the mortgage lending. From 2009 onwards, the amount on loans in the scores bellow 620 is zero (Table IV), meaning that low scores' borrowers were firmly contained since then. By that year, lending moved markedly to the higher scores (see the shape of the bars moving between years 2008 and 2009, in Table III and Table IV). This effect is also captured in the score stability index that jumps from 0,10 to 0,28 in 2009 (Table I, row 5). As a consequence there is an increase in the concentration by scores from 14% in 2008 to more than 20% since 2009 (Table I, row 4). Although this seems to be a reasonable prudential measure, we draw attention towards potential excessive lending bias and concentration in the highest scores, which requires a more precise risk-based pricing in this score bands. The number and amount of loans diminished after 2009.

TABLE I: MAIN INDICATORS FOR MORTGAGE LOANS ORIGINATED BETWEEN 1ST JANUARY 1999 AND THE 1ST QUARTER OF 2013. SOURCE: FREDDIE MAC

Indicator	Origination year														
	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013 (Q1)
Total loans (thousands)	1.095	787	1.757	1.685	1.930	1.131	1.324	1.083	1.069	986	1.513	788	556	787	247
Total original amount (billion US $)	138	104	260	262	311	188	240	202	202	210	345	177	131	192	58
Avg original loan amount ('000 US $)	126	132	148	156	161	167	181	187	189	213	228	224	236	244	237
Scores concentration index[6]	13%	13%	13%	13%	14%	13%	13%	12%	12%	14%	20%	19%	20%	21%	20%
Scores stability index[7]	n.a.	0,02	0,01	0,00	0,03	0,02	0,03	0,00	0,00	0,10	0,28	0,00	0,00	0,01	0,00
Average interest rate (AIR) (%)[8]	7,31	8,18	6,58	6,58	5,78	5,86	5,88	6,44	6,41	6,10	5,02	4,81	4,59	3,81	3,64
AIR - FIX 30 (%)	-0,13	0,13	-0,39	0,04	-0,05	0,02	0,01	0,03	0,07	0,07	-0,02	0,12	0,14	0,15	…
AIR at a low score - FIX 30 (%)[9]	0,02	0,33	-0,14	0,29	0,12	0,16	0,14	0,17	0,25	0,49	0,44	0,56	0,51	0,46	…
AIR at the highest scores - FIX 30 (%)[9]	-0,17	0,04	-0,50	-0,07	-0,09	-0,02	-0,04	-0,02	-0,02	-0,04	-0,07	0,05	0,07	0,12	…

…: not available.

TABLE II: AVERAGE ORIGINAL RATE OF THE LOAN BY SCORE. LOANS ORIGINATED IN THE PERIOD 1999-2013(Q1). UNIT: %. SOURCE: FREDDIE MAC

Score bucket	Credit risk assessment	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013 (Q1)
Unknown	Unpublished	7,49	8,28	6,74	6,74	5,91	5,98	5,92	6,30	6,28	6,40	5,31	5,10	5,14	3,97	3,91
[300;550[	Highest risk	7,50	8,29	6,88	6,88	5,97	6,04	6,09	6,75	6,92	6,77	5,48	5,46	…	3,71	3,62
[550;575[		7,50	8,35	6,93	6,94	6,02	6,10	6,14	6,72	6,81	6,75	5,62	5,44	5,38	4,00	…
[575;600[		7,53	8,53	7,00	7,00	6,07	6,10	6,11	6,65	6,67	6,60	5,56	5,29	4,99	3,95	…
[600;625[		7,46	8,38	6,83	6,83	5,95	6,00	6,01	6,58	6,59	6,52	5,48	5,25	4,96	4,12	3,92
[625;650[		7,40	8,29	6,73	6,73	5,89	5,95	5,97	6,53	6,53	6,48	5,44	5,22	4,97	4,12	3,95
[650;675[		7,36	8,23	6,65	6,65	5,84	5,90	5,93	6,50	6,48	6,36	5,34	5,14	4,90	4,07	3,90
[675;700[		7,33	8,20	6,59	6,60	5,81	5,88	5,90	6,46	6,44	6,20	5,22	5,01	4,80	3,96	3,82
[700;725[		7,30	8,16	6,56	6,56	5,78	5,86	5,88	6,44	6,41	6,13	5,11	4,89	4,69	3,87	3,71
[725;750[		7,27	8,13	6,53	6,53	5,75	5,84	5,86	6,43	6,38	6,06	5,04	4,82	4,62	3,81	3,65
[750;775[		7,25	8,10	6,50	6,50	5,73	5,81	5,83	6,40	6,35	6,02	5,00	4,78	4,57	3,79	3,62
[775;800[		7,26	8,08	6,47	6,47	5,72	5,81	5,81	6,37	6,31	5,98	4,97	4,76	4,54	3,77	3,60
[800;850]	Lowest risk	7,27	8,09	6,47	6,47	5,74	5,82	5,83	6,39	6,32	5,99	4,97	4,74	4,52	3,78	3,60
Weighed average rate		7,31	8,18	6,58	6,58	5,78	5,86	5,88	6,44	6,41	6,10	5,02	4,81	4,59	3,81	3,64

… : not applicable.

[5] A rule of thumb for the subprime mortgage is a loan of a borrower with a score inferior to 620. Some lenders also consider a subprime mortgage if the borrower has a score as high as 680 and the down payment is less than 5% of the loan.

[6] We used Herfindahl-Hirschman Index (HHI), for which values below 20% are commonly considered acceptable.
[7] We used population stability index, for which values bellow 0,25 are commonly considered normal.
[8] Calculated as the weighted average of rates by score buckets.
[9] For low scores we considered the scores in the range [600; 625]; for the highest scores we considered the scores in the range [800; 850].

TABLE III: TOTAL NUMBER OF LOANS BY SCORE IN THE YEAR. LOANS ORIGINATED IN THE PERIOD 1999-2013(Q1). SOURCE: FREDDIE MAC

Score bucket	Credit risk	Origination year															Entire period
		1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013 (Q1)	
Unknown	Unpublished	8	10	10	6	1	1	1	1	1	1	0	0	0	0	0	40
[300;550[	Highest risk	3	4	7	6	3	2	2	1	2	1	0	0	0	0	0	30
[550;575[	↑	5	5	10	9	4	3	3	3	4	2	0	0	0	0	0	47
[575;600[		13	12	25	27	15	14	15	15	17	6	1	0	0	0	0	160
[600;625[		40	32	67	63	50	40	43	39	41	16	3	1	1	1	0	437
[625;650[		84	61	127	112	103	78	85	75	74	38	13	8	5	5	2	869
[650;675[		122	85	183	168	178	123	133	111	110	65	30	18	13	13	5	1.355
[675;700[		151	103	227	213	242	151	170	135	134	102	75	43	31	37	13	1.828
[700;725[		171	112	257	239	275	160	183	149	143	130	146	79	53	70	23	2.190
[725;750[		195	125	278	262	313	172	186	146	139	143	210	106	74	104	34	2.487
[750;775[		193	135	309	303	379	197	214	167	162	182	328	162	116	166	53	3.066
[775;800[		98	89	225	240	315	160	212	172	170	211	473	240	172	250	76	3.103
[800; 850[	Lowest risk	11	13	32	37	52	30	77	71	74	91	235	130	91	142	41	1.125
Global		1.095	787	1.757	1.685	1.930	1.131	1.324	1.083	1.069	986	1.513	788	556	787	247	16.737
Contribution of the year		7%	5%	10%	10%	12%	7%	8%	6%	6%	6%	9%	5%	3%	5%	1%	100%

TABLE IV: TOTAL AMOUNT GRANTED BY SCORE IN THE YEAR. LOANS ORIGINATED IN THE PERIOD 1999-2013(Q1). UNIT: US B$. SOURCE: FREDDIE MAC

Score bucket	Credit risk	Origination year															Entire period
		1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013 (Q1)	
Unknown	Unpublished	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	6
[300;550[	Highest risk	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	4
[550;575[	↑	1	1	1	1	1	0	0	0	1	0	0	0	0	0	0	7
[575;600[		2	2	3	4	2	2	3	2	3	1	0	0	0	0	0	24
[600;625[		5	4	9	9	8	6	7	7	7	3	0	0	0	0	0	66
[625;650[		10	8	18	17	16	13	15	13	13	7	2	1	1	1	0	136
[650;675[		15	11	27	26	29	20	24	20	20	12	6	3	3	3	1	219
[675;700[		19	14	34	33	39	25	31	25	25	21	16	9	7	8	3	309
[700;725[		22	15	39	38	45	27	34	28	28	28	33	17	12	17	5	388
[725;750[		25	17	42	42	52	29	35	28	27	31	48	24	17	25	8	452
[750;775[		25	18	47	49	63	34	40	32	32	41	78	38	29	42	13	580
[775;800[		12	11	32	37	49	26	39	33	34	47	112	56	43	64	19	615
[800; 850[	Lowest risk	1	1	4	5	7	4	13	12	13	18	49	27	20	32	9	215
Total original UPB ($B)		138	104	260	262	311	188	240	202	202	210	345	177	131	192	58	3.020
Contribution of the year		5%	3%	9%	9%	10%	6%	8%	7%	7%	7%	11%	6%	4%	6%	2%	100%

TABLE V: CUMULATIVE DEFAULT 1 YEARS AFTER THE LOAN WERE ORIGINATED. UNIT: AS % OF THE LOANS IN THE SCORE BUCKET

Score bucket	Credit risk assessment	Origination year													
		1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Unknown	Unpublished	1,20%	1,63%	1,33%	2,36%	2,25%	2,75%	1,22%	1,43%	1,82%	4,22%	2,08%	0,00%	0,00%	0,00%
[300;550[	Highest risk	2,96%	4,62%	6,02%	6,83%	2,72%	3,80%	4,01%	5,92%	10,67%	12,07%	6,82%	7,14%	0,00%	0,00%
[550;575[	↑	1,73%	3,23%	3,47%	4,06%	1,99%	2,68%	2,16%	3,31%	7,10%	10,72%	2,96%	3,13%	0,00%	0,00%
[575;600[		1,39%	3,23%	3,68%	3,50%	1,66%	1,65%	1,86%	2,44%	4,40%	8,49%	2,74%	1,59%	0,00%	0,00%
[600;625[		1,05%	2,42%	2,32%	2,26%	1,12%	1,17%	1,35%	1,89%	3,37%	6,37%	1,93%	1,01%	0,71%	0,56%
[625;650[		0,75%	1,42%	1,31%	1,29%	0,65%	0,77%	0,89%	1,16%	2,52%	5,22%	1,91%	0,78%	0,67%	0,32%
[650;675[		0,46%	0,80%	0,70%	0,70%	0,39%	0,44%	0,60%	0,72%	1,63%	3,37%	0,91%	0,54%	0,39%	0,20%
[675;700[		0,24%	0,42%	0,37%	0,37%	0,21%	0,24%	0,34%	0,40%	1,21%	2,25%	0,53%	0,21%	0,25%	0,13%
[700;725[		0,13%	0,28%	0,21%	0,22%	0,12%	0,16%	0,24%	0,28%	0,87%	1,57%	0,26%	0,12%	0,12%	0,07%
[725;750[		0,09%	0,16%	0,13%	0,14%	0,07%	0,09%	0,16%	0,18%	0,60%	0,98%	0,16%	0,08%	0,07%	0,04%
[750;775[		0,06%	0,10%	0,09%	0,09%	0,04%	0,06%	0,10%	0,10%	0,33%	0,54%	0,08%	0,06%	0,04%	0,03%
[775;800[		0,06%	0,08%	0,07%	0,07%	0,03%	0,05%	0,08%	0,06%	0,17%	0,25%	0,03%	0,02%	0,03%	0,02%
[800; 850[	Lowest risk	0,07%	0,11%	0,09%	0,09%	0,05%	0,05%	0,07%	0,05%	0,12%	0,16%	0,03%	0,02%	0,02%	0,01%
Global		0,28%	0,56%	0,49%	0,48%	0,19%	0,27%	0,32%	0,41%	1,02%	1,35%	0,15%	0,08%	0,07%	0,04%

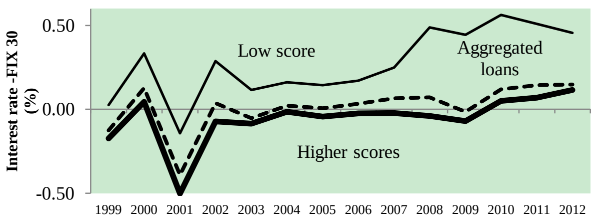
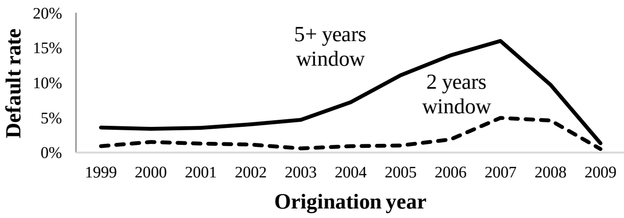


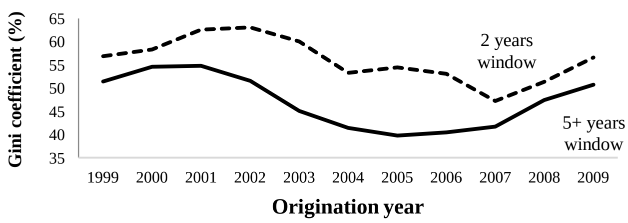
Fig. 3. Different between the average interest rate in the score bucket and the average FIX in the year.

C. Risk Assessment over Time - Default Rates, Default Misalignment and Scores Performance

Fig. 4(a) shows the cumulative default for the aggregated loans in each origination year, measuring the performance of credit 2 years after the loans have been originated (Fig. 4(a), dashed line) and 5 or more years after (Fig. 4(a), solid line). An interesting finding is that, although the mortgage bubble had expanded between 2001 and 2005, the higher loans' default rates occurred for the loans originated in the period 2004-2008. Loans originated in the beginning of the boom do not have higher defaults than the loans in the precedent years; only the borrowers that entered after 2003 had, in fact, higher risks. This finding suggests that the first borrowers of the boom were just the "lucky ones" who borrowed cheaper; only then, the opportunistic (most likely fraud) and deluded borrowers entered in the "game". Future research would much benefit from distinguishing default originated in fraud, possibly by depicting the early-stage delinquencies. Our results also reveal that, in relation to the year before Crisis, in 2006, default rates more than doubled in the loans originated in 2007, from 0.41% to 1.02% and tripled by 2008, reaching 1.35% at the aggregate level (Table V). This could be used as a benchmark for risk volatility when stress-testing current and future credit scoring and risk-based pricing models. Results also confirm that borrowers with the worst scores are more vulnerable to stressed conditions, e.g. unemployment and sudden credit-cuts, which intuition also suggests.



(a) Cumulative default rate in the loans' aggregate by origination year.



(b) Discriminatory power of scores, measured with Gini coefficient.

Fig. 4. Performance measures: the top figure shows the cumulative default by origination year and the bottom figure shows the discriminatory power of the scores. Measures are calculated in two windows: 2 years and 5+ years after the credit was originated.

By 2003, credit scoring started to decrease the ability to discriminate between good and bad customers in the short-term, (Fig. 4(b), dashed line) and by 2002 in the long-term (Fig. 4(b), solid line). This effect is likely related to population drifts, which entail further investigation of the influence of this event in credit assessment.

Finally, our research shows that real default rates by score are extremely irregular over time (Table V), which requires further consideration in models alignment, either when they are used in credit decision making or in risk-based pricing.

VI. Conclusions

Financial industry turned over-dependent on credit scoring in the advanced economies. A high proportion of the loan applications are automatically decided. In this framework, credit score is the central, if not the unique, indicator of the borrowers' credit risk.

We found evidence that the 1st year cumulative default of the borrowers in the highest scores has tripled in first years of crisis, suggesting that credit risk may has been underpriced in these cases. Two years after the crash, lending decision threshold changed and lending moved markedly to borrowers with higher scores, which led to an increase in concentration of lending in these individuals. Although this is a reasonable prudential measure, excessive lending bias and concentration towards the highest scores require more precise default estimation to correctly price credit risk.

So, credit scoring models should properly adapt to time-changing conditions and lending dynamics, to that they faithfully support risk taking and pricing. Any misalignment between the PD's by score and the real default over time will guide to inconsistent decisions and suboptimal prices. There is a new emphasis on running predictive models with the ability of sensing themselves and learn adaptively [11], [17]. This is one area where more sophistication is needed and more effort should be put to promote their wider acceptance.

References

[1] Freddie Mac, "Single Family Loan-Level Dataset," 2013.

[2] FHFA, "Conservatorship Strategic Plan: Performance Goals for 2013," ed, 2013.

[3] Freddie Mac, "Single Family Loan-Level Dataset - Summary Statistics," Freddie MacJune 2014.

[4] S. Anderson and J. Jozwik, "Building a Credit Model Using GSE Loan-Level Data," *The Journal of Structured Finance*, 2014.

[5] L. S. Goodman, *et al.*, "A Look at Freddie Mac's Loan-Level Credit Performance Data," *The Journal of Structured Finance*, vol. 19, pp. 52-61, 2014.

[6] K. M. Bianco, *The subprime lending crisis: Causes and effects of the mortgage meltdown*: CCH, Wolters Kluwer Law & Business, 2008.

[7] L. C. Thomas, *et al.*, *Credit Scoring and Its Applications*. Philadelphia: Society for Industrial and Applied Mathematics, 2002.

[8] R. Anderson, *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*: OUP Oxford, 2007.

[9] L. C. Thomas, *Consumer Credit Models: Pricing, Profit and Portfolios: Pricing, Profit and Portfolios*: Oxford University Press, 2009.

[10] L. Einav, *et al.*, "The impact of credit scoring on consumer lending," *The RAND Journal of Economics*, vol. 44, pp. 249-274, 2013.

[11] J. Gama, *Knowledge discovery from data streams*. London: Chapman & Hall/CRC, 2010.

[12] EBA, "Report on the comparability of supervisory rules and practices," 17 Dec 2013.

[13] Freddie Mac, "Single Family Loan-Level Dataset - Release Notes," ed: Freddie Mac, June 2014.

[14] Freddie Mac, "Single Family Loan-Level Dataset General User Guide", F. Mac, Ed., ed: Freddie Mac, December 2013.

[15] M. Rezac and F. Rezac, "How to Measure the Quality of Credit Scoring Models," *Finance a Uver: Czech Journal of Economics & Finance*, vol. 61, pp. 486-507, 2011.
[16] J. C. Hull, "The credit crunch of 2007: What went wrong? Why? What lessons can be learned?," ed, 2008.
[17] N. M. Adams, *et al.*, "Temporally-Adaptive Linear Classification for Handling Population Drift in Credit Scoring," in *Proceedings of COMPSTAT'2010*, ed: Physica-Verlag HD, 2010, pp. 167-176.

Maria Rocha Sousa was born in Penafiel, Portugal, in 1980. She received her MSc degree in Mathematical Engineering, in 2007, and her BSc in Mathematics, in 2003, from the University of Porto, Portugal. She is currently a PhD student in Finance, at the University of Porto, Portugal.
She is the Coordinator of the Analytics' team in the Credit Department of a major Portuguese bank, being responsible for monitoring, developing, and implementing models in the areas of credit risk and profit-driven decision-making. Her main interests are Credit Risk Dynamic Modeling, Stress-testing, Optimization and Price Modeling, in the context of International Financial Regulation.

João Gama was born in Lamego, Portugal, in 1954. He received his Ph.D. degree in Computer Science from the University of Porto, Portugal.
He is Associate Professor at the University of Porto and a senior researcher at LIAAD Inesc Tec. His main interests are Machine Learning, Data Mining, mainly in the context of data streams. He published more than 200 papers in major International conferences and journals, served as chair at ECML05, DS09, ADMA09, IDA11, and ECMLPKDD 2015. He co-organized a series of workshops on learning from data streams in conjunction with ECML-PKDD, KDD, SAC and ICML. He is member of the editorial board of MLJ, DAMI, NGC, KAIS, and PAI and he is author of a recent book in Knowledge Discovery from Data Streams.

Elísio Brandão was born in Arouca, Portugal, in 1953. He received his PhD degree in Finance, in 1988, and his MSc in Finance, in 1983, from the University of Paris IX-Dauphine, France, and his BSc in Economics, in 1978, from the School of Economics and Management (FEP) from the University of Porto.
He is Full Professor of Finance, since 2003, in School of Economics and Management and Porto Business School (PBS) at the University of Porto. His main research interests are Corporate Finance, Restructuring, M&A, Default, Credit risk and Taxation.