Expert Systems With Applications xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A new dynamic modeling framework for credit risk assessment

Maria Rocha Sousa [a,*], João Gama [a,b], Elísio Brandão [a]

[a] School of Economics and Management, University of Porto, Portugal
[b] LIAAD-INESC TEC, Portugal

ARTICLE INFO

Keywords:
Credit risk modeling
Credit scoring
Dynamic modeling
Temporal degradation
Default concept drift
Memory

ABSTRACT

We propose a new dynamic modeling framework for credit risk assessment that extends the prevailing credit scoring models built upon historical data static settings. The driving idea mimics the principle of films, by composing the model with a sequence of snapshots, rather than a single photograph. In doing so, the dynamic modeling consists of sequential learning from the new incoming data. A key contribution is provided by the insight that different amounts of memory can be explored concurrently. Memory refers to the amount of historic data being used for estimation. This is important in the credit risk area, which often seems to undergo shocks. During a shock, limited memory is important. Other times, a larger memory has merit. An application to a real-world financial dataset of credit cards from a financial institution in Brazil illustrates our methodology, which is able to consistently outperform the static modeling schema.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

In banking, credit risk assessment often relies on credit scoring models, so called PD models (Probability of Default models).[1] These models output a score that translates the probability of a given entity, a private individual or a company, becoming a defaulter in a future period. Nowadays, PD models are at the core of the banking business, in credit decision-making, in price settlement, and to determine the cost of capital. Moreover, central banks and international regulation have dramatically evolved to a setting where the use of these models is favored, to achieve soundness standards for credit risk valuation in the banking system.

Since 2004, with the worldwide implementation of regulations issued by the Basel Committee on Banking Supervision within Basel II Accord, banks were encouraged to strengthen their internal models frameworks for reaching the A-IRB (Advanced Internal Rating Based) accreditation (BCBS, 2006; BIS, 2004). To achieve this certification, banks had to demonstrate that they were capable of accurately evaluating their risks, complying with Basel II requirements, by using their internal risk models' systems, and keep their soundness. Banks owning A-IRB accreditation gained an advantage over the others, because they were allowed to use lower coefficients to weight the exposure of credit at risk, the risk weighted assets, and benefit from lower capital requirements. A lot of improvements have been made in the existing rating frameworks, extending the use of data mining tools and artificial intelligence. Yet, this may have been bounded by a certain unwillingness to accept less intuitive algorithms or models going beyond standard solutions being implemented in the banking industry, settled in-house or delivered through analytics providers.

Developing and implementing a credit scoring model can be time and resource consuming, easily ranging from 9 to 18 months, from data extraction until deployment. Hence, it is not rare that banks use unchanged credit scoring models for several years. Bearing in mind that models are built using a sample file frequently comprising 2 or more years of historical data, in the best case scenario, data used in the models are shifted 3 years away from the point they will be used. Should conditions remain unchanged, then this would not significantly affect the accuracy of the models, otherwise, their performance can greatly deteriorate over time. The recent financial crisis confirmed that financial environment greatly fluctuates, in an unexpected manner, posing renewed attention regarding models built upon time-frames that are by far outdated. By 2007–2008, many financial institutions were using stale credit scoring models built with historical data of the early-decade. The degradation of stationary credit scoring models is an issue with empirical evidence in the literature (Avery, Calem, & Canner, 2004; Crook, Thomas, & Hamilton, 1992; Lucas, 2004; Sousa, Gama, & Gonçalves, 2013b), however research is still lacking more realistic solutions.

Dominant approaches rely on static learning models. However, as the economic conditions evolve in the economic cycle, either deteriorating or improving, also varies the behavior of an individual, and his

* Corresponding author. Tel.: +351967139811.
E-mail addresses: 100427011@fep.up.pt, jsc@inescporto.pt (M.R. Sousa), jgama@fep.up.pt (J. Gama), ebrandao@fep.up.pt (E. Brandão).

[1] Other names can be used to refer to PD models, namely: credit scoring, credit risk models, scorecards, credit scorecards, rating systems, or rating models, although some have different meanings.

http://dx.doi.org/10.1016/j.eswa.2015.09.055
0957-4174/© 2015 Elsevier Ltd. All rights reserved.

UNCORRECTED PROOF

ability to repay his debt. Furthermore, the 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. Hence, default needs to be regarded as time changing.

Traditional systems that are one-shot, fixed memory-based, trained from fixed training sets, and static settings are not prepared to process the evolving data. And so, they are not able to continuously maintain an output model consistent with the actual state of environment, or to quickly react to changes (Gama, 2010). These are some of the features of classic approaches that evidence the constraints of the existing credit scoring systems. As the processes underlying credit risk are not strictly stationary, consumers' behavior and default can change over time in unpredictable ways. A few limitations to the existing approaches, idealized in the classical supervised classification paradigm, can be traced in published literature:

- The static models usually fail to adapt when the population changes. Static and predefined sample settings often lead to an incomplete examination of the dynamics influencing the problem (Gama, 2010; Hand, 2006).
- Certain assumptions that are implicit to the methods, often fail in real-world environments (Yang, 2007). These assumptions relate to:
  - *Representativeness* - the standard credit scoring models rely on supervised classification methods that run on 2-years-old static samples, in order to determine which individuals are likely to default in a future fixed period, 1 year for PD models (Thomas, 2010; Thomas, Edelman, & Crook, 2002). Such samples are supposed to be representative of the potential borrowers consumers of the future, the through-the-door population. They should also be sufficiently diverse to reflect different types of repayment behavior. However, a wide range of research is conducted in samples that are not representative.
  - *Stability and non-bias* - the distribution from which the design points and the new points is the same; classes are perfectly defined, and definitions will not change. Not infrequently there are selective biases over time. Simple examples of this occurrence can be observed when a bank launches a new product or promotes a brand new segment of customers. It can also occur when macroeconomics shifts abruptly from an expansion to a recession phase, or vice versa.
  - *Misclassification costs* - these methods assume that the costs of misclassification are accurately known, but in practice they are not.
- The methods that are most widely used in the banking industry, logistic regression and discriminant analysis are associated with some instability with high-dimensional data and small sample size. Other limitations regard to intensive variable selection effort and incapability of efficiently handling non-linear features (Yang, 2007).
- Static models are usually focused in assessing the specific risk of applicants and obligors. However, a complete picture can only be achieved by looking at the return alongside risk, which requires the use of dynamic rather than static models (Bellotti & Crook, 2013).

There is a new emphasis on running predictive models with the ability of sensing themselves and learn adaptively (Gama, 2010). Advances on the concepts for knowledge discovery from data streams suggest alternative perspectives to identify, understand and efficiently manage dynamics of behavior in consumer credit in changing ubiquitous environments. In a world where the events are not preordained and little is certain, what we do in the present affects how events unfold in unexpected ways. So far, no comprehensive set of research to deal with time changing default had much impact into practice. In credit risk assessment, a great deal of sophistication is needed to introduce economic factors and market conditions into current risk-assessment systems (Thomas, 2010).

The study presented in this paper is a large extension of a previous research that delivered the winning model within the BRICS 2013 competition in data mining and finance (Sousa, Gama, Brandão et al., 2013a; Sousa et al., 2013b). This competition opened to academics and practitioners, was focused on the development of a credit risk assessment model, tilting between the robustness of a static modeling sample and the performance degradation over time, potentially caused by market gradual changes along few years of business operation. Participants were encouraged to use any modeling technique, under a temporal degradation or concept drift perspective. In the research attached to the winning model, Sousa, Gama, and Gonçalves (2013b) have proposed a two-stage model for dealing with the temporal degradation of credit scoring models, which produced motivating results in a 1-year horizon. The winners first developed a credit scoring method using a set of supervised learning methods, and then calibrated the output, based on a projection of the evolution in the default. This adjustment considered both the evolution of the default and the evolution of macroeconomic factors, echoing potential changes in the population of the model, in the economy, or in the market. In so doing, resulting adjusted scores translated a combination of the customers' specific risk with systemic risk. The winning team (Sousa, Gama, & Gonçalves) concluded that the performance of the models did not significantly differ among classification models, like logistic regression (LR), AdaBoost, and Generalized Additive Models (GAM). However, after training in several windows lengths, they observed that the model based on the longest window has produced the best performing model over the long-run, among all competitors. This finding allowed to realize that some specifics of the credit portfolios and macroeconomic environments may reveal quite stable along time. For those cases, a model built with a static learning setting may seem appropriate, if tested during stable phases. The question yet to be answered was in which conditions credit risk models degrade? And when so, if there is any alternative modeling technique to the prevailing credit scoring models? The aim of this study is to find a clearer understanding on which type of modeling framework allows a rapid adaptation to changes, and in which circumstances a static learning setting still delivers well-performing models. With this in view, we implemented a dynamical modeling framework and two types of windows for model training, which enable testing our research questions: (a) In which conditions can a dynamic modeling outperform a static model?; (b) Is the recent information more relevant to improve forecasting accuracy?; (c) Does older information always improve forecasting accuracy?

This paper introduces a new dynamic modeling framework for credit risk assessment, imported from the emerging techniques of concept drift adaptation, in streaming data mining and artificial intelligence. The proposed model is able to produce more robust predictions in stable conditions, but also in the presence of changes, while the prevailing methods cannot. This is a promissory tool both to academics and practitioners, because unlike the traditional models, it has the ability of adjusting the predictions in the presence of changes, like inversions in the economic cycles, major crisis, or intrinsic behavioral circumstances (e.g. divorce, unemployment and financial distress). Besides the goal of enhancing the prediction of default in credit, the new modeling framework also enables developing a more comprehensive understanding of the evolution of the credit rating systems over time and anticipating unexpected events. Furthermore, we study the implications to credit risk assessment of keeping a long-term memory, and forgetting older examples, which have not been done so far.

UNCORRECTED PROOF

Please cite this article as: M.R. Sousa et al., A new dynamic modeling framework for credit risk assessment, Expert Systems With Applications (2015), http://dx.doi.org/10.1016/j.eswa.2015.09.055

Few authors have explicitly tried a dynamic modeling framework in credit risk assessment, or connected concepts. Based on a national sample of a credit reporting agency, Avery et al. (2004) show that traditional modeling often fails to consider situational circumstances, such as local economic conditions and individual trigger events, affecting the ability of scoring systems to accurately quantify individuals' credit risk. We can trace the few existing contributions in this arena over the most recent years. Sun and Li (2011) formally define financial distress concept drift and build a dynamic modeling based on instance selection. Saberi et al. (2013) worked on the concept of granularity for selecting the optimum size of the testing and training groups with a sample of credit cards of a bank operating in German. Pavlidis, Tasoulis, Adams, and Hand (2012) proposed a methodology for the classification of credit applications with the potential of adapting to population drifts.

This paper follows in Section 2 with a brief description of the main settings and concepts of the supervised learning problem and score formulation. It also presents an overview of the methods typically used in supervised learning, and specifically in credit score modeling. In Section 3, we introduce the topic of concept drift in credit default and some adaptation methods that can be promising for dynamic modeling credit risk. In Section 4 we present a case study, where we employ a set of these adaptation methods to a real-world financial dataset. First, we characterize the database and provide some intuition on the background of the problem. Then, we explain the methodology of this research. Section 5 provides the fundamental experimental results. Conclusions and future applications of the new dynamic modeling framework are traced in Section 6.

2. Settings and concepts

In this work we import some of the emerging techniques in concept drift adaptation into credit risk assessment models. This is a field of research that has been receiving much attention in machine learning over the last decade, as an answer for suitably shaping the models and processes to a reality that is ever-changing over contexts and time. The settings and definitions adopted in this paper replicate the general nomenclature surveyed by Gama, Žliobaitė, Bifet, Pechenizkiy, and Bouchachia (2014).

2.1. Supervised learning problem

Credit risk assessment can be addressed as a classification problem, a subset of supervised learning. The aim is to predict the default $y \in \{\text{good, bad}\}$, given a set of input characteristics $\mathbf{x}$. The term attribute refers to each of the possible values that a characteristic can assume; the term bin denotes a set of attributes or an interval of values in a continuous characteristics; the term example, or record, is used to refer to one pair of $(\mathbf{x}, y)$. Supervised learning classification methods try to determine a function that best separates the individuals in each of the classes, good and bad, in the space of the problem.

The model building is carried on a set of training examples - training set – collected from the past history of credit, for which both $\mathbf{x}$ and y are known. The best separation function can be achieved with a classification method. These methods include, among others, well-known classification algorithms such as decision trees (DT), support vector machines (SVM), artificial neural networks (ANN), and Generalized Additive Models (GAM). Hands-on software packages are available to the user for example in R, SAS, Matlab, and Model Builder for Predictive Analytics. In credit scoring models, the accuracy of such functions is typically assessed in separate sets of known examples – validation or out-of-sample data sets. The idea behind this procedure is to mimic the accuracy of that function in future predictions of new examples where $\mathbf{x}$ is known, but y is not.

According to the Bayesian Decision Theory (Duda, Hart, & Stork, 2001), a classification can be described by the prior probabilities of the classes $p(y)$ and the class conditional probability density function $p(\mathbf{x}|y)$ for the two classes, good (G) and bad (B). The classification decision is made according to the posterior probabilities of the two classes, which for class B can be represented as:

$$p(B|\mathbf{x}) = p(\mathbf{x}|B)\,p(B)/p(\mathbf{x}) \tag{1}$$

where $p(\mathbf{x}) = p(B)p(\mathbf{x}|B) + p(G)p(\mathbf{x}|G)$. Here, it is assumed that the costs for misclassifying a bad customer are the same as for the opposite situation, the equal costs assumption. It is worth recalling that, in real-world financial environments, the costs of failing the prediction in a real bad are by far superior to failing in a real good. In the first case, there is essentially a loss of the exposure at default, Loss Given Default (LGD), possibly mitigated with collateral. The second case affects the business, as it translates into a loss of margin. Sousa and da Costa (2008) show several possibilities to overcome this practical issue, by adapting the output of standard classification methods under the equal costs assumption to imbalanced costs for misclassification, associated with the decision and prediction tasks. It is worth discussing the related issue of class imbalance in credit scoring datasets. Quite often, these datasets contain a much smaller number of observations in the class of defaulters than in that of the good payers (Brown & Mues, 2012; Marqués, García, & Sánchez, 2012a). A large class imbalance is therefore present which some techniques may not be able to successfully handle. Baseline methods to handle class imbalance include oversampling the minority class or under sampling the majority class; Tomek links is an example of the former, SMOTE of the latter (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). Another established approach to correct imbalance adopt a cost sensitive classifier with the misclassification cost of the minority class greater than that of the majority class. Within this approach, it is worth mentioning MetaCost, a general method for making classifiers cost-sensitive (Domingos, 1999). All these methodologies, implicitly or explicitly, optimize the decision process for a specific business objective. In other words, the optimization is made for a specific trade-off between the error committed in identifying someone as defaulter when one is in fact a non-defaulter individual and the opposite type of error of diagnosing someone as non-defaulter when one is in fact a defaulter. This individualization is unconnected with our study and any of these methods can be incorporated in the methodology under research.

2.2. 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, despite the model can be used to score them. For simplicity, in this paper we will consider the problem of two classes, although the proposed methodology can easily be adapted to the other case.

The output 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, $Pr\{\text{bad}|\text{score based on } X\}$, is:

$$p(B|s(\mathbf{x})) = p(B|s(\mathbf{x}), \mathbf{x}) = p(B|\mathbf{x}), \quad \forall \mathbf{x} \in X \tag{2}$$

UNCORRECTED PROOF

Please cite this article as: M.R. Sousa et al., A new dynamic modeling framework for credit risk assessment, Expert Systems With Applications (2015), http://dx.doi.org/10.1016/j.eswa.2015.09.055

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

$$p(G|s(\mathbf{x})) = P(G|\mathbf{x}) = 1 - p(B|\mathbf{x}), \quad \forall \mathbf{x} \in X \tag{3}$$

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

$$s(\mathbf{x}) = \ln \frac{p(G|\mathbf{x})}{p(B|\mathbf{x})}, \quad \text{and } p(G|\mathbf{x}) + p(B|\mathbf{x}) = 1. \tag{4}$$

In so saying, the score may vary from $-\infty$, when $P(G|\mathbf{x}) = 0$, to $+\infty$, when $P(G|\mathbf{x}) = 1$, i.e. $s(\mathbf{x}) \in \mathbb{R}$. The probability of the default event can be written in terms of the score:

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

The most 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. In so saying, we assume that independently of the method used to determine the best separation between the two classes, good and bad, and then 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. Notwithstanding, we provide some intuitions on the technical material to survey. The basics of credit scoring and the most common approaches to build a scorecard, are further detailed in the operational research literature (Anderson, 2007; Crook, Edelman, & Thomas, 2007; McNab & Wynn, 2000; Thomas, 2009; Thomas et al., 2002). Recent advances in the area also deliver methods to build risk based pricing models (Thomas, 2009) and methodologies towards the optimization of the profitability to the lenders (Einav, Jenkins, & Levin, 2013).

2.3. Supervised classification methods

The first approach to differentiate between groups took place in Fisher's original work in (1936) for general classification problems of varieties of plants. The objective was to find the best separation between two groups, searching for the best combination of variables such that the groups were separated the most in the subspace. Durand (1941) brought this methodology to finance for distinguishing between good and bad consumer loans.

Discriminant analysis was the first method used to develop credit scoring systems. Altman (1968) introduced it in the prediction of corporate bankruptcy. First applications in retail banking were mainly focused on credit granting in two categories of loans: consumer loans, and commercial loans (for an early review and critique on the use of discriminant analysis in credit scoring see Eisenbeis (1978)). The boom of credit cards demanded the automation of the credit decision task and the use of better credit scoring systems, which were doable due to the growth of computing power. The value of credit scoring became noticed and it was recognized as a much better predictor than any other judgmental scheme. Logistic regression (Steenackers & Goovaerts, 1989) and linear programming (see Chen, Zhong, Liao, and Li, 2013 for a review) were introduced in credit scoring, and they turned out to be the most used in financial industry (Anderson, 2007; Crook et al., 2007). The use of artificial intelligence techniques imported from statistical learning theory, such as classification trees (Breiman, Friedman, Olshen, & Stone, 1984; Quinlan, 1986) and neural networks (Desai, Crook, & Overstreet Jr, 1996; Jensen, 1992; Malhotra & Malhotra, 2002; West, 2000) have arisen in credit scoring systems. Support Vector Machine (SVM) is another method based in optimization and statistical learning, that received increased attention over the last decade in research in finance, either to build credit scoring systems for consumer finance or to predict bankruptcy (Li, Shiue, & Huang, 2006; Min & Lee, 2005; Wang, Wang, & Lai, 2005). Genetic algorithms (Chen & Huang, 2003; Ong, Huang, & Tzeng, 2005), colony optimization (Martens et al., 2007), and regression and multivariate adaptive regression splines (Lee & Chen, 2005) have also been tried. Evolutionary computing (Marqués, García, & Sánchez, 2013), including genetic algorithms (Chen & Huang, 2003; Ong et al., 2005) and colony optimization (Martens et al., 2007), was also considered for credit scoring. Regression (Lee & Chen, 2005) and clustering (Wei, Yun-Zhong, & Ming-shu, 2014) techniques have also been tailored to the problem.

The choice of a learning algorithm is a difficult problem and it is often based on which happen to be available, or best known to the user (Jain, Duin, & Mao, 2000). The number of learning algorithms is vast. Many frameworks, adaptations to real-life problems, intertwining of base algorithms were, and continue to be, proposed in the literature, ranging from statistical approaches to state-of-the-art machine learning algorithms, from parametric models to non-parametric procedures (Abdou & Pointon, 2011; Baesens et al., 2003). As an alternative to using a single method, a trend that is still evolving relates to the use of hybrid systems (Hsieh, 2005; Lee, Chiu, Lu, & Chen, 2002), and ensemble of classifiers with which the outputs are achieved by a predefined sequence or rule, or a voting scheme (Marqués, García, & Sánchez, 2012b; Wang, Hao, Ma, & Jiang, 2011).

New concepts for adapting to changes (Adams, Tasoulis, Anagnostopoulos, & Hand, 2010; Pavlidis et al., 2012; Sousa et al., 2013b; Yang, 2007) and modeling the dynamics (Crook & Bellotti, 2010; Saberi et al., 2013) in populations start being exploited in credit risk assessment.

3. Dynamic modeling for credit default

3.1. Concept drift in credit default

Credit default is mostly a consequence of financial distress. A person, or a company, is in financial distress when is experiencing individual financial constraints or is being exposed to external disturbances. In private individuals, financial constraints may result from abrupt or intrinsic circumstances. In the first case, distress is usually an outcome of sorrowful events like unemployment, pay cuts, divorce, and disease. The second is most commonly related to overexposure, low assets, erratic behavior, or bad management performance. In this paper we tackle the phenomenon of concept drift in credit default, which we now briefly explain.

In the existing literature, concept drift is generally used to describe changes in the target concept, which are activated by transformations in the hidden context (Schlimmer & Granger Jr, 1986; Widmer & Kubat, 1996) in dynamically changing and non-stationary environments. As a result of these transformations, the target concept can shift suddenly or just cause a change in the underlying data distribution to the model. This means that with time, optimal features may drift significantly from their original configuration or simply lose their ability to explain the target concept. For example, a reduction of the minimum LTV (loan to value), tighten the space of possible values, which is noticed with a change in the distribution, and eventually in the credit default concept. When such drifts happen, the robustness of the model may significantly decrease, and in some situations it may no longer be acceptable.

Some authors distinguish real concept drift from virtual drift (Gama et al., 2014; Sun & Li, 2011; Tsymbal, 2004). The former refers to changes in the conditional distribution of the output (i.e., target variable) given the input features, while the distribution of the input may remain unchanged. The later refers to gradual changes in the underlying data distribution with new sample data flowing, whereas the target concept does not change (Sun & Li, 2011).

Real concept drift refers to changes in $p(y|\mathbf{x})$, and it happens when the target concept of credit default evolves in time. Such changes can occur either with or without a change in $p(\mathbf{x})$. This type of drift may happen directly as a result of new rules for defining the target classes,

UNCORRECTED PROOF

Please cite this article as: M.R. Sousa et al., A new dynamic modeling framework for credit risk assessment, Expert Systems With Applications (2015), http://dx.doi.org/10.1016/j.eswa.2015.09.055

good or bad, as those settled by regulators, when new criteria for default are demanded to the banks. Examples of these include the guidelines for the minimum number of days past due or in the materiality threshold for the amount of credit in arrears, issued with the previous Basel II Accord. Another understanding of the real concept drift in credit default is associated with indirect changes in the hidden context. In this case, credit default changes when evolving from one stage of delinquency to another. For example, most of the people with credit until five days past due tend to pay before the following installment, as most of them are just delayers. Yet, the part of debtors in arrears that also fail the next installment are most likely to be in financial distress, possibly as a result of an abrupt or intrinsic circumstance, and therefore they require more care from the bank. When arrears exceed three installments, the debtor is most certainly with serious financial constraints, and is likely to fail his credit obligations. More extreme delays commonly translate into hard stages of credit default, which require intensive tracking labor or legal actions.

Virtual drifts happen when there are changes in the distribution of the new sample data flowing without affecting the posterior probability of the target classes, $p(y|\mathbf{x})$. With time, virtual drifts may move to real concept drifts. Other interpretations can also be found in literature, for describing an incomplete representation of the data (Widmer & Kubat, 1993), and changes in the data distribution leading to changes in the decision boundary (Tsymbal, 2004). According to some authors, other events can also be seen as virtual drifts, like sampling shift (Salganicoff, 1997), temporary drifts (Lazarescu, Venkatesh, & Bui, 2004), and feature change (Salganicoff, 1997). As an example of virtual drift, we might consider the credit decision-making along the recent financial crisis. The lenders had to anticipate if a borrower would enter in default in the future (i.e. being bad). Although the macroeconomic factors have worsened, employed people with lower debt to income remained good for the lenders, and so they continued to have access to credit.

Although we are mostly interested to track and detect changes in the real target concept, $p(y|\mathbf{x})$, the methodology introduced in this paper attempts to cover both real concept and virtual drifts applied to the default concept drift detection and model rebuilding.

3.2. Methods for adaptation

Traditional methods for building a scorecard consider a static learning setting. In so doing, this task is based in learning in a predefined sample of past examples and then used to predict an actual or a potential borrower, in the future. This is an offline learning procedure, because the whole training data set must be available when building the model. The model can only be used for predicting, after the training is completed, and then it is not re-trained alongside with its utilization. In other words, when the best separation function is achieved for a set of examples of the past, it is not updated for a while, possibly for years, independently of the changes in the hidden context or in the surrounding environment. New perspectives on model building arise together with the possibility of learning online. The driving idea is to process new incoming data sequentially, so that the model may be continuously updated.

One of the most intuitive ideas for handling concept drift by instance selection is to keep rebuilding the model from a window that moves over the latest batches and use the learn model for prediction on the immediate future. This idea assumes that the latest instances are the most relevant for prediction and that they contain the information of the current concept (Klinkenberg, 2004). A framework connected with this idea consists in collecting the new incoming data for sequential batches in predefined time intervals, e.g. year by year, month by month, or every day. The accumulation of these batches generates a panel data flow for dynamic modeling.

In Finance, it remains unclear whether it is best having a long memory or forgetting old events. If on the one hand, a long memory is desirable because it allows recalling a wide range of different occurrences, in the other, many of those occurrences may no longer adjust to the present situation. A rapid adaptation to changes is achieved with a short window, because it reflects the current distribution of default more accurately. However, for the contrary reason, the performance of models built upon shorter windows worsens in stable periods. In credit risk assessment modeling, this matter has been indirectly discussed by practitioners and researchers when trying to figure the pros and cons of using a through-the-cycle (TTC) or point-in-time (PIT) schema to calibrate the output of the scorecards to the current phase of the economic cycle. For years, a PIT schema was the only option, because banks did not have sufficient historical data series. Since the implementation of the Basel II Accord worldwide, banks are required to store the data of default for a minimum 7-years period and consider a minimum of 5-years period for calibrating the scorecards.

An original idea of Widmer and Kubat (1996) uses a sliding window of fixed length with a data processing structure first-in-first-out (FIFO). Each window may consist of a single or multiple sequential batches, instead of single instances. At each new time step, the model is updated following two processes. In the first process, the model is rebuilt based on the training data set of the most recent window. Then, a forgetting process discards the data that move out of the fixed-length window.

Incremental algorithms (Widmer & Kubat, 1996) are a less extreme hybrid approach that allows updating the prediction of models to the new contexts. They are able to process examples batch-by-batch, or one-by-one, and update the prediction model after each batch, or after each example. Incremental models may rely on random previous examples or in representative selected sets of examples, called incremental algorithms with partial memory (Maloof & Michalski, 2004). The challenge is to select an appropriate window size.

4. Case study

This research evolves from a one-dimensional analysis, where we come across the financial outlook underlying the problem, to a multidimensional analysis along several points in time. The former, described in Sections 4.1, 4.2, and 4.3, is tailored to gain intuition on the default predictors and the main factors ruling the context of the problem. The latter, in Section 4.4, is designed to gradually develop and test a new dynamic framework to model credit risk.

4.1. Dataset and validation environment

The research summarized here was conducted in a real-life financial dataset, comprising 762,966 records, from a financial institution in Brazil along two years of operation, from 2009 to 2010. Each entity in the modeling dataset is assigned to a delinquency outcome - good or bad. In this problem, a person is assigned to the bad class if she had a payment in delay for 60 or more days, along the first year after the credit has been granted. The delinquency rate in the modeling dataset is 27.3%, which is in line with the high default rates in credit cards in Brazil, one of the countries with the highest default rates in the product. The full list of variables in the original data set is available in the BRICS 2013 official website. It contains 39 variables, categorized in Table 1, and one target variable with values 1 identifying a record in the bad class and 0 for the good class.

4.2. Data analysis and cleansing

Some important aspects of the datasets were considered, because they can influence the performance of the models. These aspects regard to:

UNCORRECTED PROOF

Please cite this article as: M.R. Sousa et al., A new dynamic modeling framework for credit risk assessment, Expert Systems With Applications (2015), http://dx.doi.org/10.1016/j.eswa.2015.09.055

Table 1
Predictive variables summary.

Type	#	Information
Numerical	6	Age, monthly income, time at current address, time at current employer, number of dependents, and number of accounts in the bank.
Treated as nominal	13	Credit card bills due day, 1st to 4th zip digit codes, home (state, city, and neighborhood), marital status, income proof type, long distance dialing code, occupation code, and type of home.
Binary	16	Address type proof, information of the mother and fathers names, input from credit bureau, phone number, bills at the home address, previous credit experience, other credit cards, tax payer and national id, messaging phone number, immediate purchase, overdraft protection agreement, lives and work in the same state, lives and work in the same city, and gender.
Date	1	Application date.
ID	3	Customer, personal reference, and branch unique identifiers.

- *Significant percent of zero or missing values* In exception to the variables 'lives and work in the same state' and 'previous credit experience', binary flags have 95% to 100% concentrated in one of the values, which turn them practically unworkable. The same occurs for the numerical variables number of dependents and number of accounts in the bank, both with more than 99% zeroes. The remaining variables were reasonably or completely populated.
- *Outliers and unreasonable values* The variable age presents 0.05% of applications assigned to customers with ages between 100 and 988 years. A small percent of values out of the standard ranges are observable in the variables credit card bills due day, monthly income and time at current employer. Unreasonable values are detected in the first semester of 2009, suggesting that the data were subjected to corrections from the second semester of 2009 onwards.
- *Unreliable and informal information* Little reliability on sociodemographic data is amplified by specific conditions in the background of this problem. This type of scorecards is usually based in verbal information that the customer provides, and in most of the cases no certification is made available. In 85% of the applications, no certification for the income was provided, and 75% do not have proof for the address type. Customers have little or no concern to provide accurate information. The financial industry is aware of this kind of limitations. However, in highly competitive environments there is little chance to amend them, while keeping in the business. Hence, other than regulatory imperatives, no player is able to efficiently overcome this kind of data limitations. As currently there are no such imperatives in Brazilian financial market, databases attached to this type of models are likely to keep lacking reliability in the near future.
- *Bias on the distributions of modeling examples* The most noticeable bias is in the variable monthly income, where values shift from one year to another, exhibited in Fig. 1. This is most likely related to increases in the minimum wages and inflation.

Slight variations are also observable in the geographical variables, which are possibly related with the geographical expansion of the institution. In the remaining characteristics, the correlation between the frequency distributions of 2009 and 2010 range from 99 to 100%, suggesting a very stable pattern during the analyzed period.

4.3. Data transformation and new characteristics

4.3.1. Data cleansing and new characteristics

We focused the data treatment on the characteristics that were reasonably or fully populated. Fields state, city, and neighborhood contain free text, and were subjected to a manual cleansing. Attributes with 100 or less records were assigned to a new class "Other". We could observe that there may be neighborhoods with the same name in different cities; and hence we concatenated these new cleansed fields, state and city, into the same characteristic.

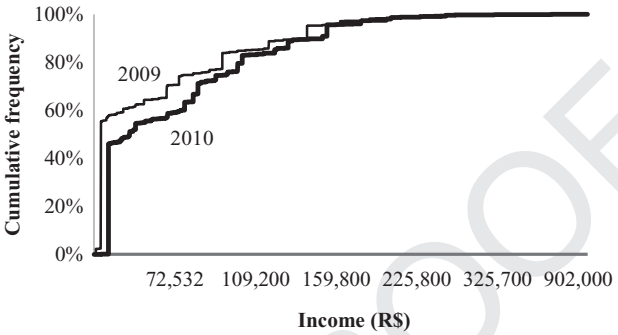


Fig. 1. Cumulative frequency of the monthly income for 2009 and 2010.

Table 2
Information values for the tested combinations.

Combination	IV
Age × income	0.315
Age × occupation	0.009
Income × marital status	0.208
Income × occupation	0.334
Income × proof of income	0.123
Age × income × occupation	0.007

4.3.2. Data transformation

Variables were transformed using the weights of evidence (WoE) in the complete modeling dataset, which is a typical measure in credit score modeling (FICO, 2006). $WoE = \ln \frac{g/G}{b/B}$, where g and b are respectively the number of good and the number of bad in the attribute, and G and B are respectively the total number of good and bad in the population sample. The larger the WoE the higher is the proportion of good customers in the bin. For the nominal and binary variables we calculated the WoE for each class. Numerical variables were firstly binned using SAS Enterprise Miner, and then manually adjusted to reflect domain knowledge. In so doing we aim to achieve a set of characteristics less exposed to overfitting. Cases where the calculation of the WoE rendered impossible - one of the classes without examples – were given an average value. The same principle was applied to values out of the expected ranges (e.g. credit card bills due day higher than 31).

4.3.3. One-dimensional analysis

The strength of each potential characteristic was measured using the information value (IV) in the period, $IV = \sum_{i=1}^{n} (g/G - b/B)WoE$, where n is the number of bins in the characteristic. The higher is the IV, the higher is the relative importance of the characteristic. In a one-dimensional basis, for the entire period, the most important characteristics are age, occupation, time at current employer, monthly income and marital status, with information values of 0.368, 0.352, 0.132, 0.117, and 0.116, respectively. Remaining characteristics have 0.084 or less.

4.3.4. Interaction terms

Using the odds in each attribute of the variables, we calculated new nonlinear characteristics using interaction terms between variables to model the joint effects. We tested six combinations, for which we present the information value in Table 2.

UNCORRECTED PROOF

4.3.5. Time series descriptive analysis

Fig. 2a shows the real concept drift along 2009–2010. The highest default rates are noticed in the first quarter of 2009, and at the end of 2010. Fig. 2b displays the evolution of the business in the same period. It exhibits two features of the business. First, we can see that the credit cards business follow an annual seasonality, increasing along each year. Second, the credit cards business is rising over time, which is related with the expansion of the branch network of the financial institution. The decrease of default rate during 2009 suggests that the decision-making process might have been slightly enhanced, when comparing to the beginning of the period.

4.4. Dynamic modeling framework

The dynamic modeling framework presented in this research considers that data is processed batch-by-batch. Sequentially, at each monthly window, a new model is learned from a previous selected window, including the most recent month. To mimic the time evolution, we assumed that the current month gradually shifts from 2009 until the third quarter of 2010.

Each learning unit for the model building was grounded on a static setting. The training of each unit consists of a supervised classification procedure, executed in three steps. First, characteristics are binned. Second, the classification model is designed with Generalized Additive Models (GAM) and a 10 fold crossed-validation, upholding the classification algorithm used to develop the winning model in the BRICS 2013 in data mining and finance (BRICS-CCI&CBIC, 2013; Sousa et al., 2013b). Concurrently, the best set of characteristics is selected until no other characteristic in the training dataset adds contribution to the information value (IV) of the model. In this application the threshold was set for a minimum increment of 0.03. Third, the performance of the model is 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 (Řezáč & Řezáč, 2011). This coefficient refers to the global quality of the credit scoring model, and ranges between −1 and 1. The perfect scoring model fully distinguishes the two target classes, good and bad, and has a Gini coefficient equal to 1. 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.

At each month, instances for modeling are selected from all previous available batches, according to a selection mechanism. We use instance selection methods to test the hypothesis under investigation. Two methods for tackling default concept drift were implemented - a full memory time window, and a fixed short memory time window with a forgetting mechanism.

The full memory time window assumes that the learning algorithm generates the model based on all previous instances (Fig. 3 a). The process is incremental, so every time a new instance arises, it is added to the training set, and a new model is build. This schema should be appropriate to detect mild concept drifts, but it is unable to rapidly adapt to major changes. Models of this schema should perform suitably in stable environments. A shortcoming of this incremental schema is that the training dataset quickly expands which may requires a huge storage capacity, and constrain the use of some classification algorithms, to be able of processing the expanding dataset.

In the fixed short memory time window, the model development uses the most recent window. With this schema, illustrated in Fig. 3 b, a new model is build in each new batch, by forgetting past examples. The fundamental assumption is that past examples have low correlation with the current default concept. Under this setting, the dynamic modeling should quickly adapt to changes. The most extreme case of short memory time window is when only the current example is considered to train the new model, which represents to the online learning without any memory of the past. A deficiency of this method is that it often lacks of generalization ability in stable conditions that is amplified with extremely short windows.

These modeling frameworks enable comparing these configurations between themselves, and also compare them with the model reached with a static learning setting. The research questions of this study should be answered following the reasoning:

- If the full memory time window outperforms the other schema, then more recent data are not fundamental for the prediction; the environment of the decision-making should be in a stable phase. Otherwise, the default concept is drifting, and so the most recent data are more relevant for the prediction.
- If a model built with static learning in the first window of the period has the best performance, then older data can improve the prediction. This may happen, for example, when a new credit product is launched, and the credit decision-making criteria are adjusted afterwards. In such case, the oldest data are more representative, as they can illustrate a more diverse range of risk behaviors. Otherwise, over the long-run, dynamic modeling should outperform the model learnt with static setting.

5. Experimental results

We assessed the performance of the sequential models built with the dynamic modeling framework introduced in the previous section, through the period 2009 and 2010. The experimental design was drawn for assessing the performance in the modeling period, in the short-term, and in the farthest-term. In each model rebuilding, the performance in the modeling period was assessed in the test set. Additionally, using two out-of-sample windows, we measured the short-term performance of the model in the month following the development, and the farthest-term performance was measured in the last quarter of 2010. Although we have considered monthly windows for developing the model, for the long-run assessment we chose a quarterly window instead of a single month. In so doing, potential atypical properties of the decision-making process at the end of the year were smoothed.

In this section, we provide further evidence on temporal degradation of static credit scoring. Then, we challenge the robustness of the new concept of dynamic modeling against a static model, developed with a traditional framework. We finally present and discuss the results for the two sliding-window configurations - full memory and short memory.

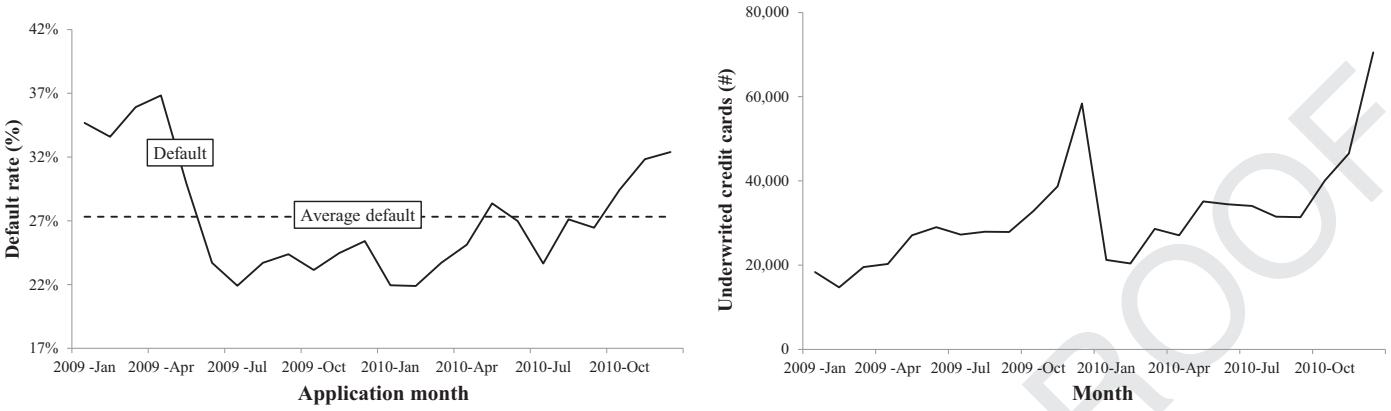
5.1. Temporal degradation of static credit scoring

The temporal degradation of the credit scoring is detected when measuring the performance of each model in the sequence generated with the dynamic modeling. Fig. 4a and b exhibit the Gini coefficient for each model, measured in the modeling test set and two different out-of-sample windows, one month after rebuild the model, and in the farthest quarter in the period (2010 Q4).

Fig. 4a shows the performance along the entire period with the short memory configuration. One month after rebuilding the model, the performance curve is always below the performance measured in the modeling period, showing that the performance consistently decreases one month after rebuilding the model. When evaluating the performance with the full memory configuration, in Fig. 4b, the extent of degradation within a month is not consistent over the period. During the first semester of 2009, performance measured in the month after rebuilding the model is slightly superior to the one measured in the modeling period, and from that point onwards, it is marginally inferior. This may suggest that the short-term

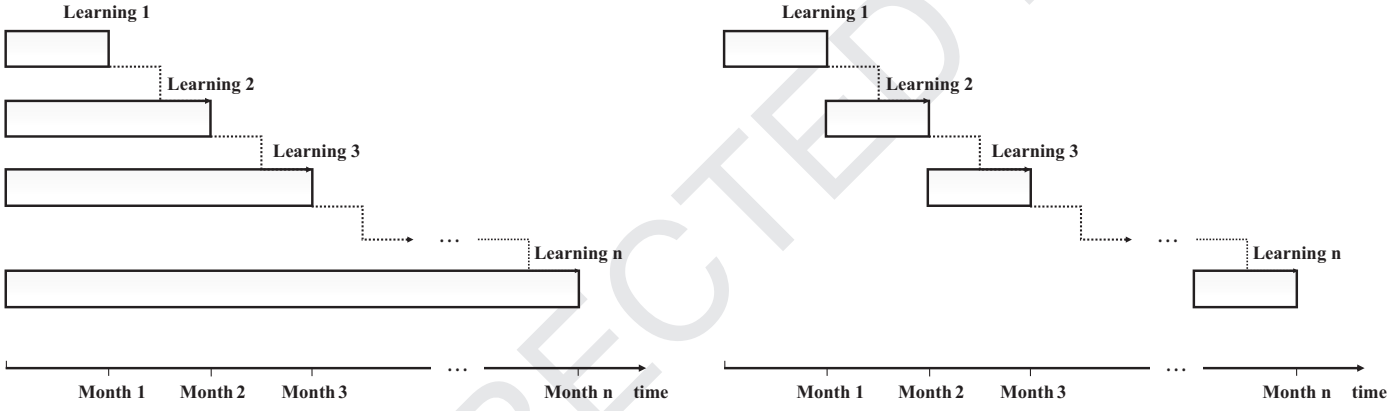
UNCORRECTED PROOF

Please cite this article as: M.R. Sousa et al., A new dynamic modeling framework for credit risk assessment, Expert Systems With Applications (2015), http://dx.doi.org/10.1016/j.eswa.2015.09.055



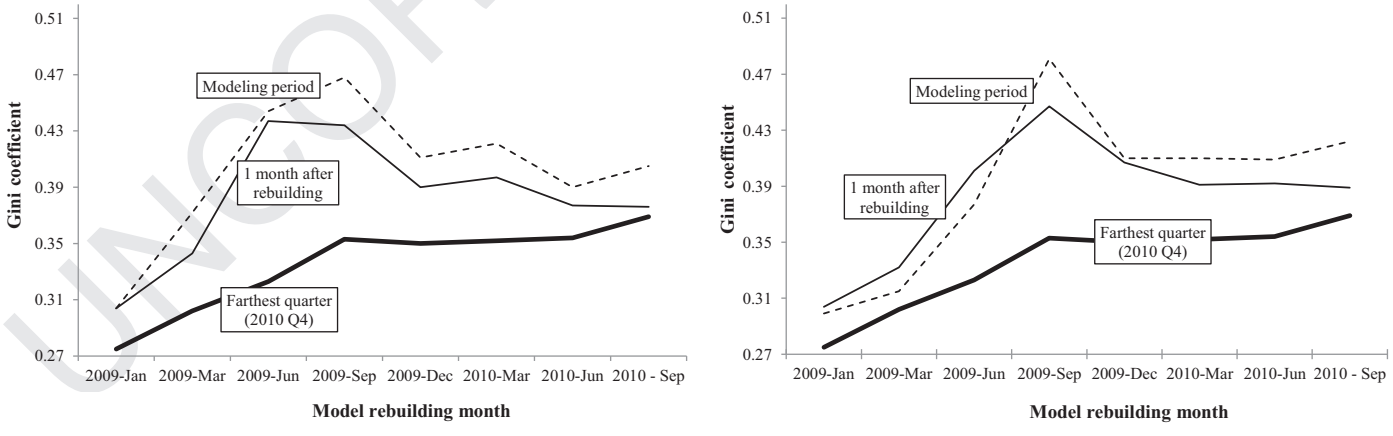
(a) Default rate. (b) Number of new credit cards contracts.

Fig. 2. Default rate and new contracts in the period 2009–2010.



(a) Full memory time window. (b) Fixed short memory time window.

Fig. 3. Configurations for tackling concept drift in credit default.



(a) Short memory. (b) Full memory.

Fig. 4. Gini coefficient of the sequence of models produced with the dynamical modeling.

performance is more similar to the performance in the modeling period when using the full memory configuration.

The extent of degradation is higher, when the performance of the model is measured at the end of the period (2014 Q4). The farthest is the point of the prediction from the point of the development; the highest is the extent of degradation of the performance. These effects are consistently perceived on the two windowing configurations - short memory (Fig. 4a) and full memory (Fig. 4b).

Considering the real performance of the models one month after they were built, the average degradation of the models sequentially constructed, shown in Table 3, is 0.02 in the short memory and 0.01 in the full memory configuration. In the farthest quarter in the period

UNCORRECTED PROOF

Please cite this article as: M.R. Sousa et al., A new dynamic modeling framework for credit risk assessment, Expert Systems With Applications (2015), http://dx.doi.org/10.1016/j.eswa.2015.09.055

Table 3
Average degradation of the sequence of models produced with the dynamic modeling.

Memory type	Gini index			Degradation	
	Modeling period	Month after rebuilding	Farthest quarter (2010 Q4)	Month after rebuilding	Farthest quarter (2010 Q4)
Short	0, 40	0, 38	0, 33	−0, 02	−0, 07
Full	0, 39	0, 38	0, 33	−0, 01	−0, 06

(2010 Q4) the degradation reaches 0.07 in the short memory configuration and 0.06 in the full memory schema.

Although degradation can be observed in all models of the sequence, updating the model always yields the best discrimination between the target classes - good and bads.

5.2. Dynamic versus static

The proposed dynamic modeling framework enables a major improvement of the initial static model, which was trained with the sample in the first month of 2009 (2009 M1).

Fig. 5a shows the immediate performance achieved with the dynamic modeling – full and short term memory – versus the static model, measured in the month following the development. Fig. 5b shows the performance of the models in each point in time, measured in the farthest quarter of the period. Consistently, for both memory configurations and performance criteria, immediate or in the farthest quarter, either the static or the dynamic modeling performances improve until the third quarter of 2009, which might reflect the enhancement of the set of characteristics $\mathbf{x}$ that was partially corrected over that period.

In Fig. 5a, we observe a certain overlap between the immediate performance achieved with the two types of memory configurations - short and full. For all the periods, the short-term performance increases until the third quarter of 2009, and slightly decreases from that point onwards. The extent of improvement with the dynamic modeling reaches 0.05 in 2010.

Fig. 5b shows that the farthest-term performance of the first model in the sequence of the dynamic modeling, same as the static model, is significantly improved with the sequential rebuilding until the third quarter of 2009, possibly as a consequence of the enhancement of the set of characteristics. In this period, performance increases from 0.28 to 0.36, meaning that the risk assessment is enhanced with the new dynamic modeling, rather than the static. From that quarter onwards, the long-run predictions given by the dynamic modeling slightly improve, and always outperform the static frame. This suggests that, the new incoming data allow a better knowledge of the new context. Although we know beforehand that the increase in performance is somewhat a consequence of the training being nearer to the out-of-sample validation window, still we can see that using the newest data improves the initial prediction given by the static model (2009 Q1).

5.3. Memory - keep or lose it

The new dynamic modeling framework enables investigating on whether it is preferable keeping a long-term memory or forgetting older observations, or if they are equivalents in some contexts. From the second semester of 2009 onwards, the best results in the farthest-term (2014 Q4) are reached with the full memory configuration. However, we realize that there is a certain overlap between the performances of the sequence of models resulting from the two types of memory configurations, both for the short-term and for the farthest-term. This suggests that, in the period, the information contained in the older examples remain appropriate for the default target, and that the context is not drifting as a result of particular changes in the set of characteristics. Hence, drifts in particular characteristics, like income, translate into virtual drifts because they did not have an impact in the distribution of target concept, $p(y|\mathbf{x})$. To some extent, the immediate performance, exhibited in Fig. 5a, decreases during 2010 from 0.44 to 0.38%, which could be interpreted as the presence of a drift. However, as the timeframe is small, it remains uncertain if it is a transitory outcome or a persistent drift in the context, potentially caused by changes in features that are not represented in the set of characteristics available for modeling in this application, like macroeconomic data.

6. Conclusions

This research presents a new modeling framework for credit risk assessment that extends the prevailing credit scoring models built upon historical data static settings. Our framework mimics the principle of films, by composing the model with a sequence of snapshots, rather than a single picture. Within the new modeling schema, predictions are made upon the sequence of temporal data, and are suitable for adapting to the occurrence of real concept drifts, translated by changes in the population, in the economy or in the market. It also enables improving the existing models based on the newest incoming data.

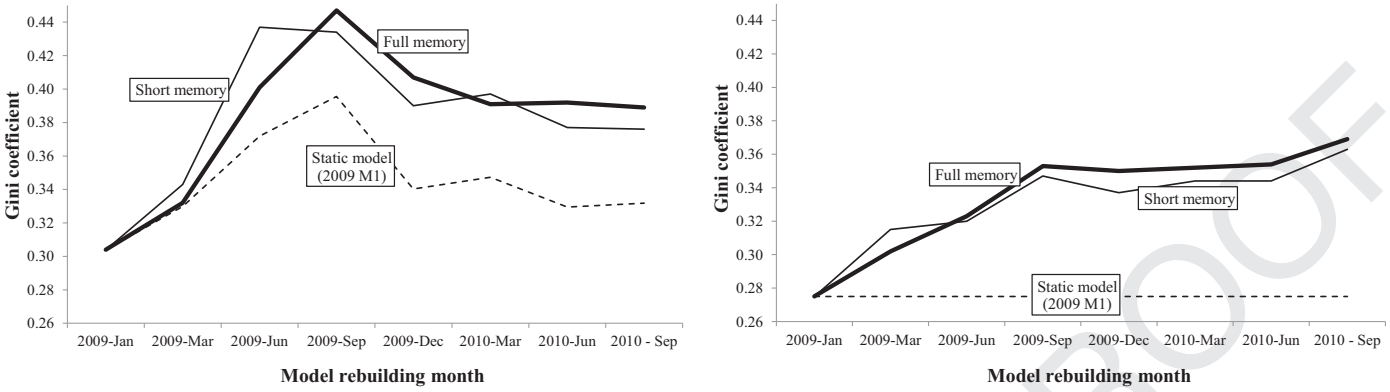
We present an empirical simulation using a real-world financial dataset of 762,966 credit cards, from a financial institution in Brazil along two years of operation. A first conclusion is that monthly updates avoid the degradation of the model following the development. Secondly, newest data consistently improve the forecasting accuracy, when compared to the previous models in the sequence of dynamic modeling, both in a short-term as in a full-term memory configuration. Particularly, the static model available at the beginning of the period is outperformed by every succeeding model, suggesting that the dynamic modeling framework has the ability of improving the prediction by integrating new incoming data. Third, a slight dominance is achieved with the full-term memory, suggesting that older information remains meaningful for predicting default target within the analyzed period.

In banking industry, prevailing credit scoring models are developed from static windows and kept unchanged possibly for years. In this setting, the two basic mechanisms of memory, short-term and long-term memory are fundamental to learning, but are still overlooked in current modeling frameworks. As a consequence, these models are insensitive to changes, like population drifts or financial distress. The usual outcomes are the default rates rising and abrupt credit cuts, as those that were observed in the U.S. in the aftermath of the last Global crisis (as documented by Sousa, Gama, and Brandão (2015)). This problem could be overcome with the proposed framework, since it would allow to gradually relearning along time and changes.

Still, there are some real business problems with rebuilding models over time. First, lenders have little incentive to enhance the existing rating systems frameworks because there is a recycling idea that it expensive and time-consuming to build new scorecards. Then, they need to be internally tested and validated, and then regulators need to approve them. Second, regulators still promote models whose coefficients do not change over time. This is one area where practice is

UNCORRECTED PROOF

Please cite this article as: M.R. Sousa et al., A new dynamic modeling framework for credit risk assessment, Expert Systems With Applications (2015), http://dx.doi.org/10.1016/j.eswa.2015.09.055



(a) Short-term performance.

(b) Farthest-term performance, 2010 Q4.

Fig. 5. Performance with the dynamic modeling – full and short memory – versus the static model (2009 M1).

far distant from the technical advances, and new thoughts, like simplifying current decision layers, need to be encouraged.

There are some important topics in default concept drift that we did not consider, which we defer for future research. While this paper provides convincing results, some additional simulations using real-world datasets from highly stressed economic environments and longer time frames would be valuable. Second, modeling the delinquency presents a specificity since a window of time is required in order to measure the outcome, i.e. the true class, before the new model is built. Therefore, for forecasting, it turns out that there will be a time gap of the same length between the values of predictor variables used in the model and the first possible forecast period in the future. Although this is not a problem of the proposed methodology, future research should bring new insights to overcome this issue, with a view on practicality. Third, some good alternatives to using windows of data blocks are encouraged, which may be based on using ensembles of the models learned in the past, possible combining the two components of memory, short-term and long-term memory, or a forgetting factor method. There is some material on this going back to Adams et al. (2010). Fourth, our empirical study considered a set of fixed predictors. Therefore, future research should consider sets of predictor of variable length. This is important for detecting concept drift, because the set predictors that are being used may be quite limited to exhibit signs of change, even if they are occurring in the environment. Finally, performance is reported in this paper, but the conditions leading to difference in performance are not explored. This is another future research direction.

References

Abdou, H. A., & 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.

Adams, N. M., Tasoulis, D. K., Anagnostopoulos, C., & Hand, D. J. (2010). Temporally-adaptive linear classification for handling population drift in credit scoring. In Y. Lechevallier, & G. Saporta (Eds.), *Proceedings of compstat'2010* (pp. 167–176). Physica-Verlag HD.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance, 23*(4), 589–609.

Anderson, R. (2007). *The credit scoring toolkit: Theory and practice for retail credit risk management and decision automation*. OUP Oxford.

Avery, R. B., Calem, P. S., & Canner, G. B. (2004). Consumer credit scoring: Do situational circumstances matter? *Journal of Banking & Finance, 28*(4), 835–856.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society, 54*(6), 627–635.

BCBS (2006). International convergence of capital measurement and capital standards: A revised framework - comprehensive version. *Bank for International Settlements*.

Bellotti, T., & Crook, J. (2013). Forecasting and stress testing credit card default using dynamic models. *International Journal of Forecasting, 29*(4), 563–574.

BIS (2004). Implementation of Basel II: Practical considerations. *Bank for International Settlements*.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Belmont, California: Wadsworth International Group.

BRICS-CCI&CBIC (2013). CI algorithms competition (CIAC): Credit risk assessment system robustness against degradation and seasonal variation. http://brics-cci.org/ci-algorithms-competition-ciac/.

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications, 39*(3), 3446– 3453. http://dx.doi.org/10.1016/j.eswa.2011.09.033.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 321–357.

Chen, D., Zhong, Y., Liao, Y., & Li, L. (2013). Review of multiple criteria and multiple constraint-level linear programming. *Procedia Computer Science, 17*(0), 158–165.

Chen, M.-C., & Huang, S.-H. (2003). Credit scoring and rejected instances reassigning through evolutionary computation techniques. *Expert Systems with Applications, 24*(4), 433–441.

Crook, J., & Bellotti, T. (2010). Time varying and dynamic models for default risk in consumer loans. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 173*(2), 283–305.

Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research, 183*(3), 1447–1465.

Crook, J. N., Thomas, L. C., & Hamilton, R. (1992). The degradation of the scorecard over the business cycle. *IMA Journal of Management Mathematics, 4*(1), 111–123.

Desai, V. S., Crook, J. N., & Overstreet Jr, G. A. (1996). A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research, 95*(1), 24–37.

Domingos, P. (1999). Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining*. In *KDD '99* (pp. 155–164). New York, NY, USA: ACM.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification*. John Wiley & Sons.

Durand, D. (1941). *Risk elements in consumer installment financing*. National Bureau of Economic Research, Inc.

Einav, L., Jenkins, M., & Levin, J. (2013). The impact of credit scoring on consumer lending. *The RAND Journal of Economics, 44*(2), 249–274.

Eisenbeis, R. A. (1978). Problems in applying discriminant analysis in credit scoring models. *Journal of Banking & Finance, 2*(3), 205–219.

FICO (2006). Introduction to scorecard for FICO model builder, .

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics, 7*(2), 179–188.

Gama, J. (2010). *Knowledge discovery from data streams*. London: Chapman & Hall/CRC.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Comput. Surv., 46*(4), 44:1–44:37. doi:10.1145/2523813.

Hand, D. J. (2006). Classifier technology and the illusion of progress. *Statistical Science, 21*(1), 30–34.

Hsieh, N.-C. (2005). Hybrid mining approach in the design of credit scoring models. *Expert Systems with Applications, 28*(4), 655–665.

Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(1), 4–37.

Jensen, H. L. (1992). Using neural networks for credit scoring. *Managerial Finance, 18*(6), 15–26.

Klinkenberg, R. (2004). Learning drifting concepts: Example selection vs. example weighting. *Intelligent data analysis, 8*(3), 281–300.

Lazarescu, M. M., Venkatesh, S., & Bui, H. H. (2004). Using multiple windows to track concept drift. *Intelligent data analysis, 8*(1), 29–59.

Lee, T.-S., & Chen, I.-F. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications, 28*(4), 743–752.

UNCORRECTED PROOF

Lee, T.-S., Chiu, C.-C., Lu, C.-J., & Chen, I.-F. (2002). Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications, 23*(3), 245–254.
Li, S.-T., Shiue, W., & Huang, M.-H. (2006). The evaluation of consumer loans using support vector machines. *Expert Systems with Applications, 30*(4), 772–782.
Lucas, A. (2004). Updating scorecards: Removing the mystique. In *Readings in credit scoring: foundations, developments, and aims* (pp. 93–109). New York: Oxford University Press.
Malhotra, R., & Malhotra, D. K. (2002). Differentiating between good credits and bad credits using neuro-fuzzy systems. *European Journal of Operational Research, 136*(1), 190–211.
Maloof, M. A., & Michalski, R. S. (2004). Incremental learning with partial instance memory. *Artificial intelligence, 154*(1), 95–126.
Marqués, A. I., García, V., & Sánchez, J. S. (2012a). On the suitability of resampling techniques for the class imbalance problem in credit scoring. *Journal of the Operational Research Society, 64*(7), 1060–1070.
Marqués, A. I., García, V., & Sánchez, J. S. (2012b). Two-level classifier ensembles for credit risk assessment. *Expert Systems with Applications, 39*(12), 10916–10922.
Marqués, A. I., García, V., & Sánchez, J. S. (2013). A literature review on the application of evolutionary computing to credit scoring. *Journal of the Operational Research Society, 64*(9), 1384–1399.
Martens, D., De Backer, M., Haesen, R., Vanthienen, J., Snoeck, M., & Baesens, B. (2007). Classification with ant colony optimization. *IEEE Transactions on Evolutionary Computation, 11*(5), 651–665.
McNab, H., & Wynn, A. (2000). *Principles and practice of consumer credit risk management*. CIB Publishing.
Min, J. H., & Lee, Y.-C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications, 28*(4), 603–614.
Ong, C.-S., Huang, J.-J., & Tzeng, G.-H. (2005). Building credit scoring models using genetic programming. *Expert Systems with Applications, 29*(1), 41–47.
Pavlidis, N., Tasoulis, D., Adams, N., & Hand, D. (2012). Adaptive consumer credit classification. *Journal of the Operational Research Society, 63*(12), 1645–1654.
Quinlan, J. R. (1986). Induction of decision trees. *Machine learning, 1*(1), 81–106.
Řezáč, M., & Řezáč, F. (2011). How to measure the quality of credit scoring models. *Finance a Uver: Czech Journal of Economics & Finance, 61*(5), 486–507.
Saberi, M., Mirtalaie, M. S., Hussain, F. K., Azadeh, A., Hussain, O. K., & Ashjari, B. (2013). A granular computing-based approach to credit scoring modeling. *Neurocomputing, 122*(0), 100–115.
Salganicoff, M. (1997). Tolerating concept and sampling shift in lazy learning using prediction error context switching. *Artificial Intelligence Review, 11*(1-5), 133–155.
Schlimmer, J. C., & Granger Jr, R. H. (1986). Incremental learning from noisy data. *Machine learning, 1*(3), 317–354.
Sousa, M. R., & da Costa, J. P. (2008). A tripartite scorecard for the pay/no pay decision-making in the retail banking industry. *Frontiers in Artificial Intelligence and Applications*, 45.
Sousa, M. R., Gama, J., & Brandão, E. (2015). Links between scores, real default and pricing: Evidence from the Freddie Mac's loan-level dataset. *Journal of Economics, Business and Management, 3*(12), 1106–1114.
Sousa, M. R., Gama, J., Brandão, E., et al. (2013a). Introducing time-changing economics into Credit Scoring. *Technical Report*. Universidade do Porto, Faculdade de Economia do Porto.
Sousa, M. R., Gama, J., & Gonçalves, M. J. S. (2013b). A two-stage model for dealing with temporal degradation of credit scoring. In *Proceedings of BRICS-CCI & CBIC*.
Steenackers, A., & Goovaerts, M. (1989). A credit scoring model for personal loans. *Insurance: Mathematics and Economics, 8*(1), 31–34.
Sun, J., & Li, H. (2011). Dynamic financial distress prediction using instance selection for the disposal of concept drift. *Expert Systems with Applications, 38*(3), 2566–2576.
Thomas, L. C. (2009). *Consumer credit models: pricing, profit and portfolios: Pricing, profit and portfolios*. Oxford University Press.
Thomas, L. C. (2010). Consumer finance: Challenges for operational research. *Journal of the Operational Research Society, 61*, 41–52.
Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). *Credit scoring and its applications*. Philadelphia: Society for Industrial and Applied Mathematics.
Tsymbal, A. (2004). *The problem of concept drift: Definitions and related work*. Dublin: Computer Science Department, Trinity College.
Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert systems with applications, 38*(1), 223–230.
Wang, Y., Wang, S., & Lai, K. (2005). A new fuzzy support vector machine to evaluate credit risk. *IEEE Transactions on Fuzzy Systems, 13*(6), 820–831.
Wei, G., Yun-Zhong, C., & Ming-shu, C. (2014). A new dynamic credit scoring model based on the objective cluster analysis. In *Practical applications of intelligent systems*. In *Advances in Intelligent Systems and Computing: 279* (pp. 579–589). Springer Berlin Heidelberg.
West, D. (2000). Neural network credit scoring models. *Computers & Operations Research, 27*(11), 1131–1152.
Widmer, G., & Kubat, M. (1993). Effective learning in dynamic environments by explicit context tracking. In *Machine learning: Ecml-93* (pp. 227–243). Springer.
Widmer, G., & Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. *Machine learning, 23*(1), 69–101.
Yang, Y. (2007). Adaptive credit scoring with kernel learning methods. *European Journal of Operational Research, 183*(3), 1521–1536.

UNCORRECTED PROOF

Please cite this article as: M.R. Sousa et al., A new dynamic modeling framework for credit risk assessment, Expert Systems With Applications (2015), http://dx.doi.org/10.1016/j.eswa.2015.09.055