






# Financial Distress Prediction in an Imbalanced Data Stream Environment

Rubens Marques Chaves<sup>1</sup>(✉) , André Luis Debiaso Rossi<sup>2</sup> ,  
and Luís Paulo Faina Garcia<sup>1</sup>(✉) 

<sup>1</sup> University of Brasília, Brasília, DF 70910-090, Brazil  
`rubens.chaves@bcb.gov.br`, `luis.garcia@unb.br`

<sup>2</sup> São Paulo State University, Itapeva, SP 18409-010, Brazil  
`andre.rossi@unesp.br`

**Abstract.** Corporate bankruptcy predictions are crucial to companies, investors, and authorities. However, most bankruptcy prediction studies have been based on stationary models, and they tend to ignore important challenges of financial distress like data non-stationarity, concept drift and data imbalance. This study proposes methods for dealing with these challenges and uses data collected from financial statements quarterly provided by companies to the Securities and Exchange Commission of Brazil (CVM). It is composed of information from 10 years (2011 to 2020), with 905 different corporations and 23,834 records with 82 indicators each. The sample majority have no financial difficulties, and only 651 companies have financial distress. The empirical experiment uses a sliding window, a history and a forgetting mechanism to avoid the degradation of the predictive model due to concept drift. The characteristics of the problem, especially the data imbalance, the performance of the models is measured through AUC,  $G_{mean}$ , and F<sub>1</sub>-Score and achieved 0.95, 0.68, and 0.58, respectively.

**Keywords:** Machine Learning · Bankruptcy · Financial Distress · Data Stream · Data Imbalance · Concept Drift · Brazil · CVM

## 1 Introduction

Nowadays, markets and companies are tightly intertwined, with a huge amount of capital flowing among market players. About 23% of the capital assets and 48% of the liability of a financial institution come from other financial institutions [9] and allow better risk and capital allocation sharing among enterprises. On the other hand, it opens the way to systemic risk, as noticed during the subprime financial crisis in 2008, which had spread globally [11]. Consequently, bankruptcy or financial distress prediction (FDP) could avoid or deal with systemic risk and diminish its consequences [33]. Moreover, it is relevant because stakeholders and the corporate owner could take action before the occurrence of bankruptcy. For

instance, it could empower owners to address the financial state of the enterprise in order to avert a bankruptcy scenario [26].

The FDP using economic-financial indicators has been extensively researched since the late 1960s [22]. Altman (1968) [4] was the first relevant work about it and used a statistical tool called Multiple Discriminant Analysis (MDA) for bankruptcy prediction, which became very popular among finance professionals. Around the 1990s, scholars started to use Artificial Intelligence (AI) and Machine Learning (ML) methods for bankruptcy prediction or FDP [10,37]. In some reviews, Alaka *et al.* (2018) [2] and Shi & Li (2019) [32] have already verified that, on average, ML models have more accuracy than statistical models.

There is a gap in the studies about FDP since most of them deal with stationary data [4,5], whereas the indicators come through a data flow and are non-stationary [31,35]. They have temporal order and timestamp associated with it. Agrahari and Singh (2021) [1] state that any data sequence with a timestamp is known as a Data Stream (DS), so FDP should be treated as a data stream problem. Additionally, in real-world applications, FDP has to deal with imbalanced classes and concept drift over time.

This study integrates two fields that are typically developed separately, the FDP and the time dimension of the data, treating it in a data stream environment. The contributions of this study are (i) a benchmark of ML classifiers for FDP in a DS environment; (ii) a benchmark of methods for data imbalance from DS; (iii) an experiment using a real-world database from the CVM; (iv) a realist scenario evaluation; (v) an impact analysis about the prediction horizon increasing.

This paper is structured as follows: Sect. 2 presents concepts to understand FDP and ML in a DS environment. Section 3 brings the reviews, surveys, and relevant studies that were the starting point of this paper. Section 4 explains the strategies used to preprocess the data, deal with concept drift, train the classifiers, and metrics to measure the performance. Section 5 brings some data and charts to illustrate the selection of the best classifier. Finally, Sect. 6 presents the conclusion and future work possibilities.

## 2 Background

Financial distress refers to a situation in which an enterprise is unable to meet its financial obligations and debt repayments. In other words, it could be defined as an inability to pay debts or preferred dividends having consequences like overdrafts, liquidation for the interests of creditors, and it may lead to a statutory bankruptcy proceeding. [4]. Some symptoms include late or missed debt payments, declining credit scores, high levels of debt, and difficulty obtaining new credit [34].

J. Sun *et al.* (2014) [34] present financial distress from two different perspectives. From a theoretical perspective, it has degrees such as mild financial distress when an enterprise faces a temporary cash-flow difficulty, and it is severe when the business fails and starts statutory bankruptcy proceedings. Additionally, it is

a dynamic changing process resulting from a continuous abnormality of business operation taking months, years, or even longer to happen [1]. The second is the empirical perspective when the enterprise faces difficulty paying debts on time and renegotiating debts with creditors.

Since the 90s, ML has been used to deal with bankruptcy prediction or financial distress identification [2, 32]. In a supervised learning problem, the goal is to learn a mapping between the input vector  $X$  and the output vector  $Y$ , given that there is a training set  $D$  of input-output pairs  $(x_i, y_i)$ . Indeed there is an unknown function  $y = f(x)$  generating each  $y_i$ . Therefore, the model training has to find a hypothesis  $h$  that approximates the function  $f$ . When the output  $y_i$  is one of a finite set, the learning problem is called classification, and if it has only two classes, it is a binary classification [29]. For example, a dataset for FDP contains healthy (negative class) and non-healthy (positive class) enterprises. Thus, it is a binary classification problem.

Nowadays, data is becoming increasingly ubiquitous [15]. Researchers have responded to this trend by developing ML algorithms for DS commonly known as incremental learning, real-time data mining, *online* learning, or DS learning [15]. Each item has an associated timestamp, and predictive models must consider items temporal order in real-time [1, 19]. When the timestamp  $t$  is considered to the supervised learning set of input-output pairs  $(x_i, y_i)$  in  $D$ , the problem is described as a set of tuples with timestamp mark  $D^t = \{(x_1^t, y_1^t), (x_2^t, y_2^t), \dots, (x_n^t, y_n^t)\}$ . Where  $i$  is a natural number bounded by  $1 \leq i \leq n$ , and identifies a element of the data chunk at the moment  $t$ .

H. M. Gomes *et al.* (2019) [15] define concept drift as a change in the statistical properties of a DS over time, and highlight that it occurs when the distribution of target concepts in a DS changes, leading to a degradation in the models' results. In the Eq. 1,  $P^t(x_i, y_i)$  is the probability of an element  $x_i$  receiving the label  $y_i$  at time  $t$ . However, over time, this probability may change. It is a common problem in DS environments, where data is constantly generated and updated, making it challenging to maintain the accuracy.

$$\exists x : P^t(x_i, y_i) \neq P^{t+1}(x_i, y_i) \quad (1)$$

In some datasets, the classes are not equally distributed, which means that at least one of them is in the minority concerning the others [13]. It biases the learning process towards the majority class and impairs the model generalization. There are two types of imbalance: intrinsic, when imbalance is something natural to the problem, for example, the financial situation of companies that are usually healthy, with a minority facing financial troubles; and extrinsic, which occurs when the imbalance results from a failure in the data collection [15].

Besides that, F. Shen *et al.* (2020) [31] have already noticed that some metrics used to evaluate ML models, such as accuracy, are not suitable for imbalanced data. It occurs when the metric uses more elements from the majority class distorting the result. Thus, it is necessary to use other set of metrics. For example, true positive rate (TPR), also known as sensitivity or recall [25], harmonic mean of precision and sensitivity when beta is equal 1 ( $F_1$ ) [25], geometric mean of

specificity and sensitivity ( $G_{mean}$ ) [25], Area Under the Curve of Receiver Operating Characteristic (AUC-ROC) [25], and Area Under the Curve of Precision and Sensitivity (AUC-PS) [30].

### 3 Related Work

The recent interest in FDP can be justified by the evolution of ML methods which has opened new possibilities and has achieved better results [5,32]. On the other hand, the academy's interest in DS learning is more recent and dates from the 2000s. The data nature is changing, the technology is collecting data all the time, and the computational power is not increasing at the same rate [14]. Given the current industry needs, there are challenges to address before the application of DS learning in real-world problems [15]. For instance, the concept drift challenge pervades different domains where the predictions are ordered by time, like bankruptcy prediction, FDP, and others [1].

The initial studies about FDP identification date from 1968 [4]. Even though it is not a new research field, in recent years, there has been a growing interest in financial and business [24,32]. T. M. Alam *et al.* (2020) [3] highlight that predicting financial distress poses two significant challenges. Firstly, the combination of economic and financial indicators, which remains a difficult task despite the efforts of specialists. Secondly, it is necessary to address the problem of data imbalance since in real-world scenarios, the amount of healthy enterprises is much larger than those facing financial distress.

S. Wanget *et al.* (2018) [36] consider two problems inherent to DS: data imbalance and concept drift. Both are very present, usually together. The authors point out that although this combination of problems frequently exists in real situations, few studies address these issues, and propose: (i) a *framework* to handle these cases; (ii) some algorithms to minimize these problems jointly. In addition, the authors highlight the lack of studies to assess the effects of data imbalance on misconceptions.

J. Sun *et al.* (2019) [35] and F. Shen *et al.* (2020) [31] noticed that previous studies on FDP seldom consider the problem of concept drift and neglect how to predict the industry financial distress in a DS environment. Both used data from Chinese companies, the sliding window method and realized that the data imbalance problem is an obvious issue related to FDP. To address it they used SMOTEBoost and Adaptive Neighbor SMOTE-Recursive Ensemble Approach (ANS-REA), respectively. J. Sun *et al.* (2019) [35] verified the existence of concept drift in FDP and associated the use of the sliding window method as the reason for outperforming stationary models. To overcome the concept drift, F. Shen *et al.* (2020) [31] used a sliding window and a forgetting mechanism. Additionally, they suggested parameter optimization and different forgetting mechanism to improve accuracy. Despite 70 attributes usage, the authors proposed the addition of new financial and non-financial indicators in the model.

This study proposes a benchmark evaluation of some ML classifiers already used for FDP in a DS environment, like Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT) [31,35], and adds

XGBoost and CatBoost commonly used in stationary environments [20, 22]. Another benchmark is about methods for data imbalance like SMOTE (Synthetic Minority Over-Sampling Technique) [8], and its variants like BorderlineSMOTE [16], ADASYN [18], SVM SMOTE [27], SMOTEENN [6] e SMOTE-Tomek [28] because its popularity [12]. The idea is to evaluate them through an experiment using a real-world database from the CVM and also evaluate the impact on the model's results after increasing the prediction horizon.

## 4 Methodology

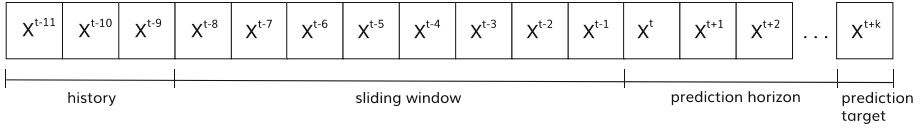
This study has gathered data from the companies listed in the CVM<sup>1</sup>. The most important documents were: the asset balance sheet, balance sheet of liabilities, income statement, and cash flow statement. They were used to produce a dataset with 23,834 entries and 82 economic-financial indicators, organized into 40 quarters over ten years (2011 to 2020). The data is strongly imbalanced: 2.73% are data of companies in financial distress situation, while 97.27% are not.

The sequence of quarters  $X^{t-h}, \dots, X^{t-2}, X^{t-1}, X^t, X^{t+1}, X^{t+2}, \dots, X^{t+k}$  is the DS where  $t$  is the present,  $t-h$  is a past moment and  $t+k$  are quarters not presented to the model yet. Each quarter  $X$  is a set of distinct data companies  $x$  with 82 attributes each. Companies in a past quarter ( $X^{t-h}$ ) have a label ( $Y^{t-h}$ ), which can be "financial distress" or "normal"; companies in the present quarter ( $X^t$ ) or ahead ( $X^{t+i}, i \in 1, \dots, k$ ) have no label and are the ones to be predicted by the model.

In this proposal, the model is trained with data from a sliding window and a subset of the historical data, as shown in Fig. 1. The *sliding window* is used to deal with concept drift and minimize its impact on the model performance. It comprises the eight most recent quarters of labeled data and its size is fixed a priori. The *history* data comprises data quarters older than those in the sliding window set and includes only instances of the minority class. These data are used to reduce the imbalanced problem, but passes through a forgetting mechanism to reduce the importance of old instances. It is an adaptation of exponential weighting scheme [23]:  $f(h) = 1 - \exp^{-\alpha h}$ , where  $h$  is the distance to the oldest quarter of the sliding window set and  $\alpha$  is a forgetting coefficient. The function  $f(h)$  returns the proportion of elements to forget for a specific historical quarter  $h$ . The *prediction target*, also known as the test set, is the data quarter which will be predicted by the model using the financial indicators, which are already known at time  $t$ . The *prediction horizon* ( $k$ ) specifies how many quarters in advance the prediction will be performed. In this work, we assume the values 2, 4, 8, 12, 16, 20, and 24 quarters.

In addition to using historical data containing only cases from the minority class, this study also applies oversampling techniques to increase the number of instances of companies in financial distress to mitigate the problem of data imbalance. The idea is to create synthetic samples to increase the minority class to 50% and 100% of the majority class to identify the best balancing rate ( $Rt$ ) for

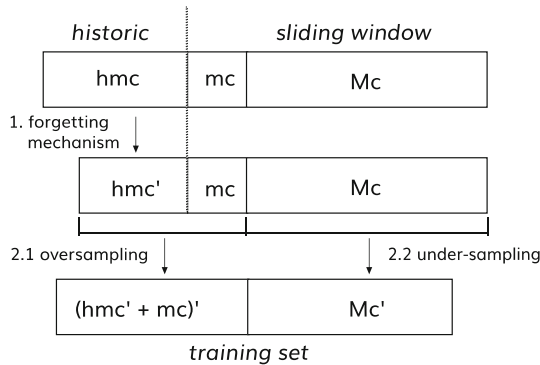
<sup>1</sup> <https://dados.cvm.gov.br/>.



**Fig. 1.** Sliding window after eleven quarter with three historic quarters and eight quarters for the window

each model using methods to balance the data (*i.e.* SMOTE, BorderlineSMOTE, ADASYN, SVMSMOTE, SMOTEENN, and SMOTETomek).

In the preprocessing phase, a set of instances of the minority class is over-sampled before model training. Figure 2 illustrates the training set generation. In step 1, it selects all instances of the minority class from the sliding window  $mc$  and merge with instances from the history after the forgetting mechanism  $hmc'$ . Hence, it merges the selected set  $hmc' + mc$  with the sliding window majority class  $Mc$ . Then, in step 2.1, it applies the oversampling technique. Step 2.2 minimizes the creation of synthetic instances by the under-sampling technique  $(hmc' + mc)' + Mc'$ .



**Fig. 2.** Data preprocessing to generate the training set

After the sliding window has accumulated enough data, with eight quarters, the training process is conducted in rounds using the prepared training set. Because of the time dependence of the data, the nested cross-validation on time series [21] is more appropriate to train and validate the classification model (*i.e.* LR, SVM, RF, DT, XGBoost and CatBoost).

Mainly because of the imbalance condition of the dataset and the importance of correct classification of the minority samples, the metrics used to evaluate models performance were  $F_1$ -Score and  $G_{mean}$ . Other important metrics were the AUC-ROC to measure the overall accuracy [7, 17] and the AUC-PS to complement the analysis [30].

**Table 1.** Classifiers results using balancing technique and balancing strategies (0, 0.5 and 1).

Metric	Preprocessing	Rt	LR	DT	SVC	RF	XGBoost	CatBoost	
F <sub>1</sub> -Score	-	-	0.0812±0.00	<i>0.3562±0.12</i>	0.0000±0.00	0.4133±0.03	0.4203±0.14	0.4581±0.14	
	SMOTE	0.5	0.2019±0.03	0.3516±0.10	0.0013±0.00	0.3976±0.03	0.5132±0.12	0.5683±0.12	
		1.0	0.2611±0.02	0.3138±0.12	0.0705±0.00	0.3640±0.03	0.5247±0.12	0.5665±0.13	
	B.SMOTE	0.5	0.1912±0.02	0.3200±0.12	0.0019±0.00	0.3658±0.03	0.4649±0.13	0.5179±0.15	
		1.0	0.2319±0.01	0.3231±0.12	0.0656±0.00	0.3442±0.03	0.4657±0.13	0.5124±0.15	
	ADASYN	0.5	0.1730±0.01	0.3187±0.14	0.0038±0.00	0.3605±0.03	0.4729±0.14	0.5241±0.15	
		1.0	0.2211±0.01	0.3143±0.13	0.0558±0.01	0.3301±0.04	0.4769±0.13	0.5165±0.16	
	SVMSMOTE	0.5	0.1885±0.03	0.3614±0.10	0.0020±0.00	<i>0.4474±0.02</i>	<i>0.5302±0.12</i>	0.5743±0.13	
		1.0	0.2524±0.02	0.3602±0.13	<i>0.0706±0.00</i>	0.4273±0.03	0.5273±0.12	0.5792±0.13	
	SMOTEENN	0.5	0.1977±0.03	0.3517±0.11	0.0016±0.00	0.4108±0.03	0.5167±0.12	0.5719±0.13	
		1.0	0.2649±0.01	0.3307±0.11	0.0671±0.00	0.3730±0.03	0.5288±0.12	<b>0.5812±0.12</b>	
	SMOTETomek	0.5	0.1896±0.03	0.3521±0.11	0.0013±0.00	0.4030±0.03	0.5142±0.12	0.5621±0.12	
		1.0	<i>0.2674±0.01</i>	0.3113±0.11	0.0705±0.00	0.3700±0.03	0.5216±0.12	0.5681±0.12	
	G <sub>mean</sub>	-	-	0.2490±0.09	<i>0.5575±0.11</i>	0.0000±0.00	0.5124±0.11	0.5272±0.14	0.5518±0.14
		SMOTE	0.5	0.5429±0.06	0.5604±0.10	0.0038±0.01	0.5089±0.14	0.6614±0.12	0.6624±0.12
			1.0	0.6837±0.12	0.5190±0.13	<i>0.2334±0.06</i>	0.4838±0.14	0.6845±0.12	0.6722±0.13
		B.SMOTE	0.5	0.5226±0.04	0.5346±0.11	0.0057±0.01	0.4744±0.15	0.6008±0.13	0.6218±0.15
			1.0	0.6320±0.13	0.5298±0.12	0.2055±0.07	0.4572±0.15	0.6103±0.13	0.6236±0.15
ADASYN		0.5	0.5147±0.03	0.5266±0.13	0.0111±0.01	0.4714±0.18	0.6171±0.14	0.6248±0.15	
		1.0	0.6474±0.13	0.5193±0.13	0.1552±0.05	0.4448±0.18	0.6325±0.13	0.6275±0.16	
SVMSMOTE		0.5	0.5238±0.06	0.5698±0.11	0.0057±0.01	<i>0.5544±0.14</i>	0.6766±0.12	0.6698±0.13	
		1.0	0.6843±0.12	0.5620±0.12	0.2316±0.06	0.5407±0.15	<b>0.6961±0.12</b>	<i>0.6882±0.13</i>	
SMOTEENN		0.5	0.5307±0.06	0.5596±0.11	0.0047±0.01	0.5218±0.14	0.6620±0.12	0.6668±0.13	
		1.0	0.6842±0.12	0.5379±0.12	0.2284±0.06	0.4907±0.14	0.6908±0.12	0.6865±0.12	
SMOTETomek		0.5	0.5228±0.06	0.5575±0.10	0.0038±0.01	0.5137±0.14	0.6618±0.12	0.6593±0.12	
		1.0	<i>0.6927±0.11</i>	0.5161±0.12	0.2321±0.06	0.4871±0.14	0.6834±0.12	0.6754±0.12	
AUC-ROC		-	-	0.7525±0.01	0.6571±0.07	0.5106±0.03	0.9102±0.03	0.9405±0.02	0.9436±0.03
		SMOTE	0.5	0.8004±0.02	0.6590±0.06	0.6035±0.07	<i>0.9298±0.03</i>	0.9379±0.03	0.9484±0.02
			1.0	0.8304±0.03	0.6397±0.07	0.6461±0.01	0.9280±0.03	0.9395±0.03	0.9514±0.02
		B.SMOTE	0.5	0.7832±0.02	0.6448±0.06	0.6027±0.03	0.9223±0.03	0.9331±0.03	0.9471±0.03
			1.0	0.8073±0.02	0.6424±0.07	0.5932±0.03	0.9223±0.03	0.9304±0.03	0.9469±0.03
	ADASYN	0.5	0.7859±0.02	0.6440±0.08	0.6279±0.05	0.9265±0.03	0.9315±0.03	0.9469±0.03	
		1.0	0.8145±0.02	0.6395±0.07	0.6051±0.02	0.9189±0.04	0.9310±0.03	0.9481±0.02	
	SVMSMOTE	0.5	0.7950±0.03	0.6655±0.06	0.5943±0.07	0.9291±0.02	0.9364±0.03	0.9474±0.03	
		1.0	0.8268±0.03	0.6620±0.07	<i>0.6520±0.01</i>	0.9251±0.03	0.9374±0.03	0.9479±0.03	
	SMOTEENN	0.5	0.7990±0.02	0.6595±0.06	0.5962±0.08	0.9269±0.03	0.9404±0.03	0.9506±0.02	
		1.0	0.8306±0.02	0.6483±0.07	0.6410±0.01	0.9287±0.03	<i>0.9424±0.03</i>	<b>0.9520±0.02</b>	
	SMOTETomek	0.5	0.7959±0.03	<i>0.6586±0.06</i>	0.5986±0.07	0.9280±0.03	0.9384±0.03	0.9485±0.02	
		1.0	<i>0.8314±0.02</i>	0.6370±0.07	0.6454±0.00	0.9276±0.03	0.9389±0.03	0.9513±0.02	
	AUC-PS	-	-	0.0768±0.00	0.3831±0.11	0.0349±0.00	0.5821±0.14	0.5677±0.14	0.6136±0.15
		SMOTE	0.5	0.1134±0.01	0.3745±0.09	0.0503±0.01	0.5733±0.13	0.5657±0.14	0.6352±0.14
			1.0	0.1363±0.01	0.3504±0.10	0.0646±0.00	0.5368±0.14	0.5768±0.13	0.6342±0.13
		B.SMOTE	0.5	0.1029±0.01	0.3454±0.11	0.0475±0.00	0.5725±0.14	0.5250±0.15	0.6067±0.15
			1.0	0.1195±0.01	0.3475±0.11	0.0617±0.02	0.5552±0.15	0.5131±0.15	0.5954±0.15
ADASYN		0.5	0.0996±0.01	0.3432±0.13	0.0481±0.00	0.5522±0.16	0.5190±0.15	0.6039±0.16	
		1.0	0.1175±0.01	0.3413±0.13	<i>0.0765±0.02</i>	0.5282±0.17	0.5083±0.14	0.5919±0.16	
SVMSMOTE		0.5	0.1078±0.02	0.3886±0.09	0.0473±0.01	<i>0.5872±0.14</i>	0.5724±0.14	0.6388±0.14	
		1.0	0.1311±0.01	<i>0.3926±0.11</i>	0.0764±0.01	0.5631±0.14	0.5816±0.14	0.6360±0.14	
SMOTEENN		0.5	0.1134±0.02	0.3764±0.10	0.0494±0.01	0.5789±0.13	0.5719±0.14	<b>0.6414±0.13</b>	
		1.0	0.1383±0.01	0.3621±0.09	0.0680±0.01	0.5438±0.13	<i>0.5822±0.14</i>	0.6386±0.13	
SMOTETomek		0.5	0.1096±0.02	0.3809±0.10	0.0498±0.01	0.5749±0.13	0.5691±0.14	0.6336±0.13	
		1.0	<i>0.1386±0.01</i>	0.3437±0.10	0.0667±0.01	0.5438±0.13	0.5787±0.13	0.6335±0.13	

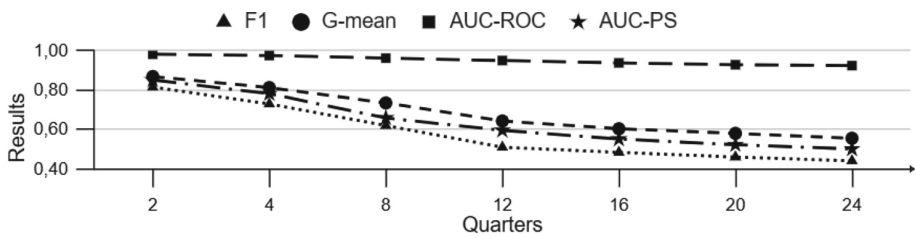
## 5 Results

Firstly the classifiers' performances were evaluated after preprocessing approach using different balancing strategies rate ( $Rt = \{0, 0.5, 1\}$ ) and different prediction horizon (2, 4, 8, 12, 16, 20, and 24 quarters), the average and the standard deviation of metrics (F<sub>1</sub>-Score, G<sub>mean</sub>, AUC-ROC and AUC-PS) were computed. In Table 1, the average best results of F<sub>1</sub>-Score, G<sub>mean</sub>, AUC-ROC and AUC-PS for the combination of classifier, preprocessing approach, and balancing strategies were presented. The italic values are the best results of a classifier among

balancing techniques (in a column), and the bold number is the best result for a specific metric among all classifiers.

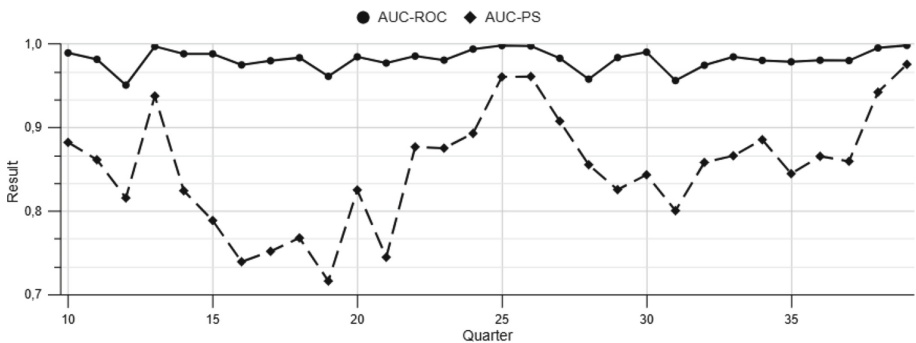
In Table 1 it is possible to observe that the best predictive performance (bold values) is related to the CatBoost for most of the metrics analysed. Additionally, the best balancing technique is the SMOTEENN with a balancing rate  $Rt = 1$  because it presents the higher value for  $F_1$  and AUC-ROC, while the values for  $G_{mean}$  and AUC-PS are very close. Hence, the CatBoost classifier and SMOTEENN with a balancing strategy of 100% are better.

Next analyses are about the impact of prediction horizon variation (2, 4, 8, 12, 16, 20 and 24) on the metrics  $F_1$ -Score,  $G_{mean}$ , AUC-ROC and AUC-PS, using the CatBoost classifier and SMOTEENN (100%) as balancing technique. In Fig. 3, the x-axis is the prediction horizon quarters and the y-axis is the average result of the metrics over time.



**Fig. 3.** The evaluation metrics  $F_1$ -Score,  $G_{mean}$ , AUC-ROC and AUC-PS after changing the precision horizon from 2 quarters to 24 quarters

Figure 3 shows that the prediction horizon and classifier performance measured by  $F_1$ -Score,  $G_{mean}$ , AUC-ROC, and AUC-PS are inversely proportional. It means that when the prediction horizon increases, the classifier performance decreases. Hence, the best classifier result is when the prediction horizon is smaller (*i.e.* 2 quarters). The AUC-ROC behavior differs from others because the strong data imbalance rate impacts it, and it should be analyzed together with AUC-PS [30].



**Fig. 4.** Classifier AUC-ROC and AUC-PS evolution during training using prediction horizon of 2 quarters



The final analysis is about the CatBoost behavior over cross-validation on time series using the SMOTEENN with a balancing strategy of 100% and a prediction horizon of 2 quarters. Figure 4 shows a chart where the x-axis is the classifier result, and the y-axis is the quarter, a variation curve of AUC-ROC, a variation curve of AUC-PS. As time goes, the AUC-ROC remains always above 0.95 while AUC-PS get its worst value (0.7164) in the 19th quarter and increases until it reaches its best value (0.9760) in the 39th quarter. Thus, there is an increasing trend for the AUC-PS curve because of the accumulation of financial distress instances in history, which reduces the number of synthetic samples necessary to balance the data chunk. The valleys in the AUC-ROC and AUC-PS curves (quarters 12, 19, 28, and 31) may be interpreted as concept drifts.

On this study the best overall results were obtained using CatBoost and the balancing method of SMOTEENN. The results may be compared with F. Shen *et al.* [31] because they used a very similar methodology and forgetting coefficient set to “1”, its best classifier is RF and the balancing method is the ANS-REA. The performance of AUC-ROC was better in this study (0.9519 vs. 0.9138), however, the  $F_1$ -Score (0,5811 vs. 0.8003) and the  $G_{mean}$  (0,6865 vs. 0.8783) was not. In this study the minority class represents 2.73% of samples while in the Shen’s study the minority class represents 33%, this reasonable difference explains the difference in the  $F_1$ -Score and the  $G_{mean}$  between the studies.

## 6 Conclusion

This study investigates the FDP with strongly imbalanced data in a DS environment combining different classifiers, preprocessing data balancing techniques, and data selection to deal with concept drift. This approach is more suitable than those that deal with stationary data because enterprises’ economic-financial indicators are susceptible to concept drift [35], and it can be the basis for building an autonomous FDP solution.

The empirical experiment uses data from 2011 to 2020, consisting of 651 financially distressed companies and 23,183 matching normal enterprises, all of which are listed on the Brazilian stock exchange from CVM. The results demonstrate that FDP in a DS environment is possible even when the data is strongly imbalanced. The use of balancing techniques improved the metrics’ results in all cases. Hence, they are import tools to deal with imbalanced data and should be added to machine learning pipelines to deal with FDP in DS. When the CatBoost is used with SMOTEENN, balancing the minority class at 100% of majority, it outperforms the best results of the classifiers LR, DT, SVC, RF, and XGBoost. In  $F_1$ -Score it is superior by 117.35%, 63.17%, 723.23%, 29.91%, and 9.62%, in AUC-ROC it is superior by 14.51%, 44.55%, 46.01%, 2.39% and 1.02%, in AUC-PS it is superior by 360.75%, 62.66%, 734.77%, 8.75%, and 9.69%. The exception is in  $G_{mean}$  because it is superior to DT, SVC, and RF by 23.44%, 194.86%, and 24.13%, although it is slightly inferior to LR and XGBoost by 0.65% and 1.13%.

Differently, from other studies about FDP in dynamic environments [31, 35] that did not use AUC-PS, in this study it complemented the information from

AUC-ROC and helped to identify the moments of concept drift and the way the model recovered from a drift. It also showed that the sliding window, the history, and the forgetting mechanism are important to deal with the concept drift. Thus, it should be used more often when dealing with imbalanced data and data streams. Additionally, the prediction horizon should be increased with caution because it severely impacts the classifiers performance.

The experiment performed during this study may be improved with the use of a period larger than ten years because this could enlarge the history and fewer synthetic instances of the minority class will be necessary. The forgetting coefficient should also be adjusted to more accurate parameter optimization to improve the accuracy because, with the current value, the mechanism forgets most historic instances till the second quarter of the history. Different sliding window lengths could be tried or even an adaptive sliding window [23] could be used. Moreover, further research could be conducted on concept drift to identify different types of drift and adapt the models after detecting a drift [1]. For this purpose, the dataset used in this study is available on GitHub<sup>2</sup>.

**Acknowledgment.** This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

## References

1. Agrahari, S., Singh, A.K.: Concept drift detection in data stream mining: a literature review. *J. King Saud Univ. Comput. Inf. Sci.* (2021). <https://doi.org/10.1016/j.jksuci.2021.11.006>
2. Alaka, H.A., et al.: Systematic review of bankruptcy prediction models: towards a framework for tool selection. *Expert Syst. Appl.* **94**, 164–184 (2018). <https://doi.org/10.1016/j.eswa.2017.10.040>
3. Alam, T.M., et al.: Corporate bankruptcy prediction: an approach towards better corporate world. *Comput. J.* **64**(11), 1731–1746 (2020). <https://doi.org/10.1093/comjnl/bxaa056>
4. Altman, E.I.: Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J. Financ.* **23**(4), 589–609 (1968). <https://doi.org/10.1111/j.1540-6261.1968.tb00843.x>
5. Barboza, F., Kimura, H., Altman, E.: Machine learning models and bankruptcy prediction. *Expert Syst. Appl.* **83**, 405–417 (2017). <https://doi.org/10.1016/j.eswa.2017.04.006>
6. Batista, G.E.A.P.A., Prati, R.C., Monard, M.C.: A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl.* **6**(1), 20–29 (2004). <https://doi.org/10.1145/1007730.1007735>
7. Bradley, A.P.: The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recogn.* **30**(7), 1145–1159 (1997). [https://doi.org/10.1016/S0031-3203\(96\)00142-2](https://doi.org/10.1016/S0031-3203(96)00142-2)
8. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**(1), 321–357 (2002). <https://doi.org/10.5555/1622407.1622416>

<sup>2</sup> <https://github.com/rubensmchaves/ml-fdp>.

9. Duarte, F., Jones, C.: Empirical network contagion for U.S. financial institutions. FRB of NY Staff Report **1**(826) (2017)
10. Efrim Boritz, J., Kennedy, D.B.: Effectiveness of neural network types for prediction of business failure. *Expert Syst. Appl.* **9**(4), 503–512 (1995). [https://doi.org/10.1016/0957-4174\(95\)00020-8](https://doi.org/10.1016/0957-4174(95)00020-8). <https://www.sciencedirect.com/science/article/pii/0957417495000208>. Expert systems in accounting, auditing, and finance
11. Eichengreen, B., Mody, A., Nedeljkovic, M., Sarno, L.: How the subprime crisis went global: evidence from bank credit default swap spreads. *J. Int. Money Financ.* **31**(5), 1299–1318 (2012). <https://doi.org/10.1016/j.jimonfin.2012.02.002>
12. Fernández, A., García, S., Herrera, F., Chawla, N.V.: Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *J. Artif. Intell. Res.* **61**(1), 863–905 (2018)
13. Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F.: Learning from Imbalanced Data Sets. Springer, Cham (2018). <https://doi.org/10.1007/978-3-319-98074-4>
14. Gama, J.: A survey on learning from data streams: current and future trends. *Progress Artif. Intell.* **1**(1), 45–55 (2012). <https://doi.org/10.1007/s13748-011-0002-6>
15. Gomes, H.M., Read, J., Bifet, A., Barddal, J.P., Gama, J.: Machine learning for streaming data: state of the art, challenges, and opportunities. *ACM SIGKDD Explor. Newsl.* **21**(2), 6–22 (2019). <https://doi.org/10.1145/3373464.3373470>
16. Han, H., Wang, W.-Y., Mao, B.-H.: Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: Huang, D.-S., Zhang, X.-P., Huang, G.-B. (eds.) *ICIC 2005*. LNCS, vol. 3644, pp. 878–887. Springer, Heidelberg (2005). [https://doi.org/10.1007/11538059\\_91](https://doi.org/10.1007/11538059_91)
17. Hanley, J., Mcneil, B.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982). <https://doi.org/10.1148/radiology.143.1.7063747>
18. He, H., Bai, Y., Garcia, E.A., Li, S.: ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1322–1328 (2008). <https://doi.org/10.1109/IJCNN.2008.4633969>
19. He, H., Chen, S., Li, K., Xu, X.: Incremental learning from stream data. *IEEE Trans. Neural Netw.* **22**(12), 1901–1914 (2011). <https://doi.org/10.1109/TNN.2011.2171713>
20. Huang, Y.P., Yen, M.F.: A new perspective of performance comparison among machine learning algorithms for financial distress prediction. *Appl. Soft Comput.* **83**, 105663 (2019). <https://doi.org/10.1016/j.asoc.2019.105663>
21. Hyndman, R.J., Athanasopoulos, G.: *Forecasting: Principles and Practice*. OTexts (2021)
22. Jabeur, S.B., Gharib, C., Mefteh-Wali, S., Arfi, W.B.: CatBoost model and artificial intelligence techniques for corporate failure prediction. *Technol. Forecast. Soc. Change* **166**, 120658 (2021). <https://doi.org/10.1016/j.techfore.2021.120658>
23. Klinkenberg, R.: Learning drifting concepts: example selection vs. example weighting. *Intell. Data Anal.* **8**(3), 281–300 (2004). <https://doi.org/10.5555/1293831.1293836>
24. Kumbure, M.M., Lohrmann, C., Luukka, P., Porras, J.: Machine learning techniques and data for stock market forecasting: a literature review. *Expert Syst. Appl.* **197**, 116659 (2022). <https://doi.org/10.1016/j.eswa.2022.116659>

25. Li, Z., Huang, W., Xiong, Y., Ren, S., Zhu, T.: Incremental learning imbalanced data streams with concept drift: the dynamic updated ensemble algorithm. *Knowl.-Based Syst.* **195**, 105694 (2020). <https://doi.org/10.1016/j.knosys.2020.105694>
26. Lin, X., Zhang, Y., Wang, S., Ji, G.: A rule-based model for bankruptcy prediction based on an improved genetic ant colony algorithm. *Math. Probl. Eng.* 753251 (2013). <https://doi.org/10.1155/2013/753251>
27. Nguyen, H.M., Cooper, E.W., Kamei, K.: Borderline over-sampling for imbalanced data classification. *Int. J. Knowl. Eng. Soft Data Paradigm* **3**(1), 4–21 (2011). <https://doi.org/10.1504/IJKESDP.2011.039875>
28. Rana, C., Chitre, N., Poyekar, B., Bide, P.: Stroke prediction using Smote-Tomek and neural network. In: 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1–5 (2021). <https://doi.org/10.1109/ICCCNT51525.2021.9579763>
29. Russell, S., Norvig, P.: *Artificial Intelligence: A Modern Approach*, 3rd edn. Prentice Hall, Hoboken (2010)
30. Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **10**, 1–21 (2015). <https://doi.org/10.1371/journal.pone.0118432>
31. Shen, F., Liu, Y., Wang, R., Zhou, W.: A dynamic financial distress forecast model with multiple forecast results under unbalanced data environment. *Knowl.-Based Syst.* **192**, 105365 (2020). <https://doi.org/10.1016/j.knosys.2019.105365>
32. Shi, Y., Li, X.: A bibliometric study on intelligent techniques of bankruptcy prediction for corporate firms. *Heliyon* **5**(12), 12 (2019). <https://doi.org/10.1016/j.heliyon.2019.e02997>
33. Silva, T.C., da Silva Alexandre, M., Tabak, B.M.: Bank lending and systemic risk: a financial-real sector network approach with feedback. *J. Financ. Stab.* **38**, 98–118 (2017). <https://doi.org/10.1016/j.jfs.2017.08.006>
34. Sun, J., Li, H., Huang, Q.H., He, K.Y.: Predicting financial distress and corporate failure: a review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. *Knowl.-Based Syst.* **57**, 41–56 (2014). <https://doi.org/10.1016/j.knosys.2013.12.006>
35. Sun, J., Zhou, M., Ai, W., Li, H.: Dynamic prediction of relative financial distress based on imbalanced data stream: from the view of one industry. *Risk Manag.* **21**(4), 215–242 (2019). <https://doi.org/10.1057/s41283-018-0047-y>
36. Wang, S., Minku, L.L., Yao, X.: A systematic study of online class imbalance learning with concept drift. *IEEE Trans. Neural Netw. Learn. Syst.* **29**(10), 4802–4821 (2018). <https://doi.org/10.1109/TNNLS.2017.2771290>
37. Wilson, R.L., Sharda, R.: Bankruptcy prediction using neural networks. *Decis. Support Syst.* **11**(5), 545–557 (1994). [https://doi.org/10.1016/0167-9236\(94\)90024-8](https://doi.org/10.1016/0167-9236(94)90024-8)