A Financial Distress Prediction using a Non-stationary Dataset

Rubens Marques Chaves[1] **, André Luis Debiaso Rossi**[2] **, Luís Paulo Faina Garcia**[1]

[1]Programa de Pós-Graduação em Informática (PPGI) – Universidade de Brasília (UnB)
Caixa Postal 4.466 – 70.910-900 – Brasília – DF – Brazil

[2]Departamento de Ciências e Tecnologia – Universidade Estadual de São Paulo (Unesp)
Rua Geraldo Alckmin, 519 – Vila N. Sra. de Fátima – 18.409-010 – Itapeva – SP – Brazil

rubens.chaves@bcb.gov.br, andre.rossi@unesp.br, luis.garcia@unb.br

Abstract. *Financial distress prediction (FDP) is crucial to companies, investors, and authorities. However, most FDP studies have been based on stationary models, disregarding important challenges present on financial distress data such as non-stationarity. Therefore, the lack of real-world datasets of economic-financial indicators organized in a timeline manner is a gap to be addressed. This study proposes a comprehensive dataset of 84 economic-financial indicators from the Brazilian Securities and Exchange Commission (CVM) organized in a non-stationary manner and validated by experiments using classification models. The results of the metrics AUC-ROC, AUC-PS, F_1-Score and G_{mean} bring evidences that the dataset is suitable for FDP.*

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 [Duarte and Jones 2017]. The intertwining allows better risk and capital allocation sharing between 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 [Eichengreen et al. 2012]. Consequently, bankruptcy or Financial Distress Prediction (FDP) could avoid or deal with systemic risk and diminish its consequences [Silva et al. 2017]. Moreover, it has great worth as it may inform the corporate owner and other stakeholders in predicting bankruptcy earlier. It could support corporate owners for effective decision-making related to the corporate financial condition and also identifies the future scopes of particular corporate in the context of long-term business operations in the market [Lin et al. 2013]. Thus, since the late 1960s, academics have been addressing this issue using statistical methods [Altman 1968]. More recently, Machine Learning (ML) techniques have demonstrated their effectiveness and have surpassed the results achieved by traditional statistical models [Barboza et al. 2017].

Corporate failure is not an abrupt event but a gradual process with distinct phases [Agarwal and Taffler 2008]. Thus, it is crucial to examine the period leading up to the bankruptcy filing when the corporation began to present some difficulties [Alam et al. 2020]. Financial distress is defined as a negative term employed to describe the financial situation of an enterprise under a stressful moment, which means it has no liquidity and is struggling to satisfy its financial obligations on time fully [Sun et al. 2014].

The FDP using economic-financial indicators has been extensively researched since the late 1960s [Frydman et al. 1985, Altman et al. 1994, Sun et al. 2011, Altman 2013, Clement 2020, Jabeur et al. 2021, Duarte and Barboza 2020, Barboza et al. 2022]. These indicators come from financial data such as balance sheets, income statements, cash flow statements, accounts receivable aging reports, and budget reports [Ross et al. 2012]. Usually, these documents are provided to the shareholders and government authorities by public authorities. Additionally, the indicators are regularly updated annual and quarterly basis [Douglas and Bates 1933, Simon 1989].

Despite the interest of academics and practitioners in the topic [Kumbure et al. 2022], the availability of datasets remains a barrier. The majority of the datasets are not publicly available [Barboza et al. 2017, Barboza et al. 2022, Bragoli et al. 2022, Zou et al. 2022, Chen et al. 2022, Pilch 2021], while just few are public [Lombardo et al. 2022, Liang et al. 2016, Zieba et al. 2016, Tang et al. 2019]. Additionally, most of these datasets have considered the data as stationary and have overlooked the time dependence aspect, which has been addressed in recent studies [Sun et al. 2019, Shen et al. 2020, Kim et al. 2022].

Because of the absence of public non-stationary datasets of financial distress enterprises, this study proposes a dataset spanning 10 years (2011 to 2020) of data of Brazilian enterprises extracted from the Open Data Portal[1] of the Brazilian Securities and Exchange Commission (CVM). The 84 attributes are commonly used by practitioners and scholars as economic-financial indicators [Altman 1968, Tomczak 2016, Barboza et al. 2017, Liang and Tsai 2020, Shen et al. 2020], and were extracted and computed from accounting files, organized by quarters and enterprises. The dataset contains indicators from 905 different enterprises, consisting of 23,834 records. As the dataset is composed of real-world data, it exhibits a strong imbalance, with 2.73% belonging to financially distressed enterprises and 97.27% belonging to non-distressed ones. In addition, we carried out experiments using classification ML techniques to predict financial distress and validate the proposed dataset.

This paper is structured as follows: Section 2 presents the main concepts about FDP and ML. Section 3 presents the reviews and research used as a starting point for this study. Section 4 explains the strategies used to gather economic-financial indicators from CVM and Section 5 validates the dataset through a empirical experiment. Section 6 presents the evaluation performance results. Finally, Section 7 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. A theoretical framework of the *cash flow* or *liquid assets* defines financial distress as a result from factors like the inability to pay debts or preferred dividends and the corresponding consequences such as overdraft of bank deposits, liquidation for interests of creditors, and even entering the statutory bankruptcy proceeding [Beaver 1966, Altman 1968]. Symptoms of financial distress include late or missed debt payments, declining credit scores, high levels of debt, and difficulty obtaining

[1] `https://dados.cvm.gov.br`

new credit. If left unchecked, financial distress can lead to bankruptcy and legal action from creditors [Sun et al. 2014].

To deal with FDP is necessary to face some challenges like strong class imbalance and non-stationary data which are very present in real-world situations, usually together [Wang et al. 2018]. A dataset is considered imbalanced when the classes are not equally distributed, resulting in at least one of them being in the minority compared to the others [Fernández et al. 2018]. It can cause learning bias towards the majority class and impair the model generalization. On the other hand, non-stationary data requires attention about changes in the statistical properties of a dataset over time, and it occurs when the distribution of target concepts within dataset changes, leading to an increase in prediction errors and a decrease in the accuracy of predictive models, also known as concept drift [Gomes et al. 2019].

Since 60s the FDP have called the attention of academics, which on that time used statistic tools to predict financial distress. In 1968, an influential paper on the prediction of corporate bankruptcy using discriminant analysis was written [Altman 1968]. It was the first of many other studies about the matter [Altman et al. 1977, Altman et al. 1994, Frydman et al. 1985]. However, after some decades and the evolution of ML models, a new research concluded that these new techniques have overcome those based on statistics [Barboza et al. 2017]. Some ML models used for that purpose were Logistic Regression (LR) [Barboza et al. 2017], Support Vector Machines [Hui and Sun 2006], Decision Tree [Zibanezhad et al. 2011], Random Forest (RF) [Alam et al. 2020], eXtreme Gradient Boosting (XGBoost) [Barboza et al. 2022], Categorical Boosting (CatBoost) [Martorano 2021] and Neural Networks [Tang et al. 2019] and others. Some of them have reached accuracy higher than 90%.

Besides that, some metrics used to evaluate ML models, such as accuracy, are not suitable for imbalanced data [Shen et al. 2020]. 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 [Li et al. 2020], harmonic mean of precision and sensitivity when $\beta = 1$ (F_1-Score) [Li et al. 2020], geometric mean of specificity and sensitivity (G_{mean}) [Li et al. 2020], Area Under the Curve of Receiver Operating Characteristic (AUC-ROC) [Li et al. 2020], and Area Under the Curve of Precision and Sensitivity (AUC-PS) [Saito and Rehmsmeier 2015].

3. Related Work

The recent growing interest in FDP has been noticed, which is justified by the advances in ML techniques over the last three decades [Shi and Li 2019]. These advances have opened up new possibilities in the field of FDP. However, only a few studies have considered the non-stationary nature of economic-financial indicators and attempted to predict financial distress using them, which might pave the way towards an autonomous solution [Sun et al. 2019, Shen et al. 2020, Kim et al. 2022].

The majority of FDP studies rely on private datasets, such as Compustat[2]: database containing fundamental financial and price data for active and inactive

[2] https://www.library.hbs.edu/find/databases/compustat

Table 1. Datasets from recent studies on FDP or bankruptcy prediction cited in this article.

Article	Source	Samples	Attr.	Period	Free	Data
[Tomczak 2016]	UCI	10,503	64	2007-2013	yes	S
[Barboza et al. 2017]	Compustat	14,198	11	1985-2013	no	S
[Succurro 2017]	Orbis	1,033,661	17	2012-2014	no	S
[Liang and Tsai 2020]	UCI	6,819	96	1999-2009	yes	S
[Shen et al. 2020]	CSMAR	4,147	70	2007-2017	no	NS
[Pilch 2021]	Orbis	53,847	33	2014-2018	no	S
[Bragoli et al. 2022]	AIDA	27,133	7	2007-2015	no	S
[Barboza et al. 2022]	Economatica	1,055	17	2000-2017	no	S
[Chen et al. 2022]	CSMAR	10,731	199	2007-2019	no	S
[Lombardo et al. 2022]	American stock market	8,262	18	1999-2018	yes	S
This study	CVM	23,834	84	2011-2020	yes	NS

publicly traded companies from the United States [Barboza et al. 2017]; Economatica[3]: dataset with stock market data from Brazil, Latim America and United States [Barboza et al. 2022]; AIDA from Bureau van Dijk[4]: contains comprehensive information on companies in Italy [Bragoli et al. 2022]; China Stock Market & Accounting Research database (CSMAR)[5]: is a comprehensive research-oriented database focusing on China Finance and Economy [Shen et al. 2020, Zou et al. 2022, Chen et al. 2022], and; Orbis[6]: database which has information on close to 450 million companies and entities across the globe [Succurro 2017, Pilch 2021]. These datasets often contain extensive information that could be organized chronologically, such as by year, semester or quarter. However, the datasets are not ready to use and need further computing to extract the attributes (economic-financial indicators).

In the opposite direction, some datasets are freely available in public or personal repositories; they contain calculated economic-financial indicators and are ready to use. However, most of them do not consider the sequential order in which the indicators were generated, thus they have stationary data. Some of them are available on ML repository of the University of California Irvine (UCI)[7], such as the Polish company dataset [Tomczak 2016] and Taiwanese company dataset [Liang and Tsai 2020], OpenML repository[8], Kaggle[9] and a bankruptcy prediction dataset for American companies in the stock market on a personal repository[10] [Lombardo et al. 2022].

Table 1 summarizes the features of datasets from studies cited in this paper, ordered by publication year. It includes information about the data *source*, number of *samples* in the dataset, number of *attributes* (Attr.) used for prediction, availability of the dataset (*free*), and the column *Data* specify if the data is organized in a stationary (S) or non-stationary (NS) manner.

---

[3] `https://economatica.com/`
[4] `https://aida.bvdinfo.com`
[5] `http://cndata1.csmar.com/`
[6] `https://www.bvdinfo.com/en-gb/our-products/data/international/orbis`
[7] `https://archive.ics.uci.edu/ml/index.php`
[8] `https://www.openml.org/`
[9] `https://www.kaggle.com/`
[10] `https://github.com/sowide/bankruptcy\_dataset`

4. Proposed Dataset

The data were gathered from the CVM's open data portal, specifically from the Quarterly Information Form[11,12]. This form includes important documents (*i.e.* asset balance sheet, balance sheet of liabilities, income statement, and cash flow statement) that are organized on an annual basis and contain raw data that needs to be processed before it can be used by ML models. These documents are required to comply with the International Financial Reporting Standards (IFRS) issued by the International Accounting Standards Board (IASB) [Comissão de Valores Monetários 2022]. They follow a reporting format where accounting information (*e.g.* assets, current assets, fixed asset and inventories) is organized by lines and includes columns such as company identification, item code, item name, and item value (Brazilian currency). The company identification separates the accounting information of each company and will be replaced with anonymized value in the final dataset. The referential date identifies the quarter when the information occurred, while the item code and name are used to identify specific item, which may vary among companies.

The creation process of the dataset have to identify the right accounting items and have to transpose them from lines to attribute columns, which are directly extracted from the data files. Table 2 presents these attributes and is organized into three columns: *Document* indicating the source of information, *Indicator* representing the information name and *Attribute* identifying each attribute with a code.

These attributes are important to compute a second set of attributes that have been used for other studies based on FDP [Altman 1968, Tomczak 2016, Barboza et al. 2017, Shen et al. 2020, Liang and Tsai 2020, Barboza et al. 2022, Chen et al. 2022, Bragoli et al. 2022]. The attributes listed in Table 3 were also listed in similar study [Shen et al. 2020]. The latter column represents the attribute name in the dataset, which includes other columns not listed in the tables, such as `ID` (a sequential value for different companies) and `QUARTER` (representing the last day of the quarter). Additionally, the target label assumes two values: 0 for non-distressed companies and 1 for distressed ones, as indicated in the `LABEL` column.

In the first set of attributes (Table 2), depending on the company's business area, there may be instances where certain accounting item information is not available in the CVM's data files. For example, a bank's balance sheet does not include inventory information. In such cases, the corresponding feature value is set to zero. In the second attribute set (Table 3), most of the attributes are ratios and during its calculus the divisor may be zero, formally expressed as $\frac{x}{0}$, since it is indeterminate, its value is set to zero. In other cases, the dividend is zero, formally expressed as $\frac{0}{x}$, resulting zero, these variables are also set to zero.

Finally, the repository and is organized into 40 quarters over a ten-year period (2011 to 2020), encompassing data from 905 corporations. This results in a total of 23,834 records and includes 84 extracted and computed indicators that have already been used by scholars for prediction. The data exhibits a strong class imbalance, with 2.73% of the records representing companies in a financial distress situation, while 97.27% represent

[11] In Portuguese "*Formulário de Informações Trimestrais (ITR)*"

[12] `https://dados.cvm.gov.br/dataset/cia\_aberta-doc-itr`

Table 2. Attributes gathers directly from the CVM's data files.

Document	Indicator	Attribute
Balance Sheet (assets)	Total assets	A1
	Current assets	A2
	Availability	A3
	Receivables	A4
	Inventory	A5
	Long-term assets	A6
	Intangible assets	A7
	Tangible assets	A8
	Fixed assets	A9
	Accumulated depreciation	A10
	Accumulated amortization	A11
	Investments	A12
Balance Sheet (liabilities)	Total liabilities	A13
	Current liabilities	A14
	Non-current liabilities	A15
	Commitments ($A13 - A14$)	A16
	Net worth ($A12 - A15$)	A17
	Share capital	A18
	Reserves	A19
	Provisions	A20
	Long term loan	A21
Income Statement	Gross income	A22
	Expenses	A23
	Net earnings	A24
	Operating expenses	A25
	Operating profit	A26
	Financial result	A27
	Financial expenses	A28
	Profit before tax	A29
	Tax expenses	A30
	Net income	A31
Cash Flow Statement	Cash flows from operating activities	A32
	Cash Flows from Investing	A33
	Cash Flows from Financing	A34
Referential Form	Outstanding shares	A35

companies that are not. It is available in GitHub[13].

5. Experiment

In this section we describe the experiments carried out to evaluate the dataset for the FDP problem while preserving the order in which the instances were generated. The idea is to read data chunks on a quarterly basis and predict whether an enterprise is in financial distress or not. The sequence of quarters forms the sequence:

$$X^{t-h}, ..., X^{t-2}, X^{t-1}, X^{t}, X^{t+1}, X^{t+2}, ..., X^{t+k}$$

where t represents the present time, $t - h$ 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 84

[13] https://github.com/rubensmchaves/ml-fdp

Table 3. Attributes calculated from more than one the CVM's data files.

Category	Indicator	Attribute
Short-term liquidity	Current ratio	A36
	Quick ratio	A37
	Cash ratio	A38
Long-term liquidity	Interest coverage ratio	A39
	Debt ratio	A40
	Tangible asset coverage ratio	A41
	Ratio of equity to debt	A42
	Ratio of commitments to tangible assets	A43
Structure of assets	Liquidity ratio	A44
	Receivable assets ratio	A45
	Fixed Asset Ratio	A46
	Ratio of stockholders' equity to fixed assets	A47
	Current debt ratio	A48
Operating capacity	Operating net profit ratio	A49
	Ratio of receivables to gross income	A50
	Ratio of inventory to income	A51
	Inventory turnover	A52
	Turnover ratio of account payable	A53
	Turnover of current assets	A54
	Ratio of fixed assets to income	A55
	Total capital turnover	A56
Profitability	Return On Assets	A57
	Ratio of net profit to total assets	A58
	Ratio of net profit to current assets	A59
	Ratio of net profit fixed assets	A60
	Return On Equity (ROE)	A61
	Operating profit ratio	A62
	Ratio of total operating cost to gross revenue	A63
	Expenses to sales Ratio	A64
	Management Expense Ratio	A65
	Financial Expense Ratio	A66
Cash	Free Cash Flow	A67
	Ratio of operating cash to net profit	A68
	Ratio of operating cash to income	A69
	Cash recovery rate	A70
	Financial leverage	A71
	Operational leverage	A72
	Combined leverage	A73
Growth capacity	Growth of capital maintenance rate	A74
	Growth of capital accumulation rate	A75
	Growth of total assets rate	A76
	Growth rate of ROE	A77
	Growth rate of net profit	A78
	Growth rate of operating profit	A79
	Growth rate of operating receipt	A80
	Growth rate of operating cost	A81
Indicators per share	Earnings per share	A82
	Net asset value per share	A83
	Net cash per share	A84

attributes each. Companies in a past quarter (X^{t-h}) have a label (Y^{t-h}), which can be 1 ("financial distress") or 0 ("normal"), companies in the present quarter (X^t) or ahead (X^{t+k}) have no label and are the ones to be predicted by the model.

Moreover, this experiment uses a sliding window and a forgetting mechanism to deal with concept drift and minimize its impact on model performance. In Figure 1, the *history* comprises quarters older than those in the sliding window set and includes only instances of the minority class. The *sliding window* (w) consists of the most recent eight quarters of data, avoiding to use too old instances to train the model. The *prediction target* also known as the test set, is the set of companies indicators used by the model to predict the companies' situation, and the *prediction horizon* (k) specifies how many quarters in advance the prediction will happen. Here, we have set the size of the sliding to eight ($w = 8$), the prediction horizon to two ($k = 2$) and the history of the minority class is unlimited.

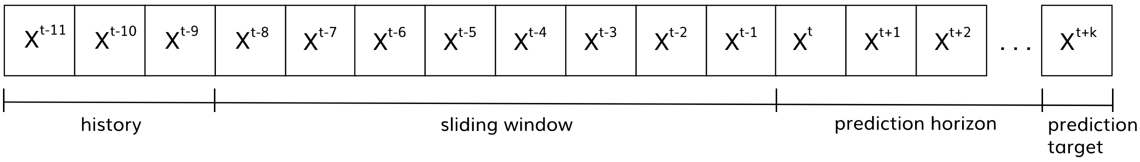

Figure 1. Sliding window after eleven quarter with three historic quarters and eight quarters for the window.

The historical data passes through a forgetting mechanism to reduce the importance of old instances. It is an adaptation of an exponential weighting scheme [Klinkenberg 2004]: $f(h) = 1 - exp^{-\alpha h}$, where h is the distance to the oldest quarter of the sliding window set and α is a forgetting coefficient. The function $f(h)$ returns the proportion of elements to forget for a specific historical quarter h.

After the sliding window has accumulated enough data (*i.e.*, eight quarters of data), the training process is conducted in rounds using the prepared training set. Because of the time dependence of the data, the nested cross-validation method for time series [Hyndman and Athanasopoulos 2021] is more appropriate to train and validate the model. The ML classifiers used for models induction were Logistic Regression (LR) [Martin 1977, Ohlson 1980], Random Forest (RF) [Breiman 2001], Decision Tree (DT) [Breiman et al. 1984], and Categorical Boost (CatBoost) [Jabeur et al. 2021] with the default values of hyperparameters (scikit-learn[14]), except the LR classifier which used the solver `liblinear` with at most 300 iterations (*i.e.* `max_iter = 300`). Additionally, given the patent class imbalance and recognizing the importance of assessing models predictive performance with respect to the minority class, we adopted the AUC-ROC [Bradley 1997, Hanley and Mcneil 1982], AUC-PS [Saito and Rehmsmeier 2015], F_1-Score [Shen et al. 2020] and G_{mean} [Shen et al. 2020] metrics. The best classifiers were selected through the analysis of the mean across 30 quarters and by using statistical tests such as Friedman and Nemenyi [Demšar 2006]. At this point, it is important to note that the random classifier for AUC-ROC is 0.5, whereas for AUC-PS, it corresponds to the imbalanced rate of 0.027 in this case [Saito and Rehmsmeier 2015].

[14] `https://scikit-learn.org/stable/index.html`

6. Results

Considering the proposed methodology, the results were generated for each classifier from the $10th$ quarter on. The mean predictive performance for the whole period are presented in Table 4 for each classifier and metric (columns). Each value is presented along with its corresponding standard deviation, which falls within an acceptable range of values. The bold values highlight the highest values per column.

Table 4. Mean values of the metrics for 30 quarters.

Classifier	AUC-ROC	AUC-PS	F_1-Score	G_{mean}
LR	0.7684±0.03	0.0790±0.03	0.0846±0.07	0.2290±0.18
DT	0.7595±0.06	0.5563±0.10	0.5427±0.10	0.7199±0.08
RF	0.9669±0.02	0.8044±0.07	0.6326±0.07	0.6847±0.05
CatBoost	**0.9826**±0.01	**0.8502**±0.05	**0.7253**±0.06	**0.7605**±0.05

The fragility of LR and DT in dealing with imbalanced data, as evidenced previously and is not apparent when looking solely at the AUC-ROC metric [Zhang et al. 2019, Cieslak and Chawla 2008]. They achieved scores of 0.76841 and 0.75948, respectively. However, this is due to the imbalanced nature of the data, which impairs the ROC curve and biases it towards the majority class. Considering the AUC-PS, F_1-Score and G_{mean}, the LR obtained 0.07902, 0.08466 and 0.22908, and DT obtained 0.55627, 0.54269 and 0.71989, respectively. In contrast, RF obtained a score of 0.80441, 0.63261 and 0.68467, and CatBoost achieved 0.85016, 0.72535 and 0.76053. When we consider AUC-PS, LR obtained a value close to that of a random classifier, while DT obtained a significantly lower value compared to RF and CatBoost, indicating their classification inefficiency (see Table 4). Between RF and CatBoost, the second obtained values slightly better than the first.

We also made an analysis of the classifiers' behavior along the quarters. The predictive performance through 30 quarters of data is shown in Figures 2, 3, 4 and 5 for AUC-ROC, AUC-PS, F_1-Score and G_{mean}, respectively. Where, the x-axis is the quarter and the y-axis is the metric result for each quarter.

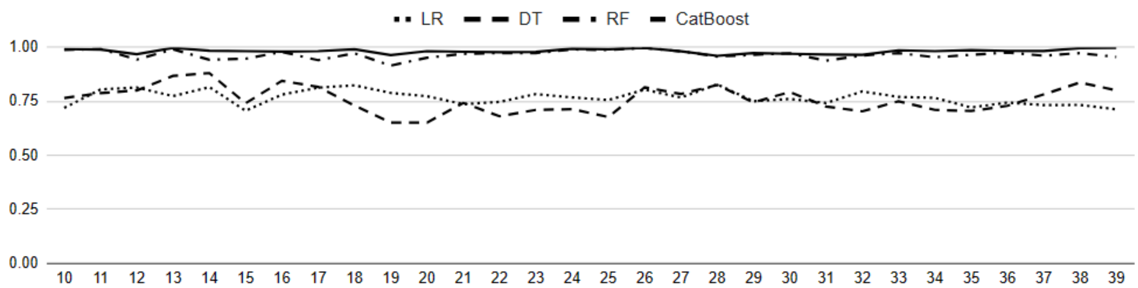

Figure 2. Classifier's AUC-ROC behavior during the period of 30 quarters (10 to 39).

Figures 2, 3, 4 and 5 show that RF and CatBoost were the most stable classifiers across the quarters, and, except for the F_1-Score, they exhibited very similar behavior. This can be observed quantitatively by examining the standard deviation values presented in Table 4. On the other hand, LR was the most unstable and significantly deviated from

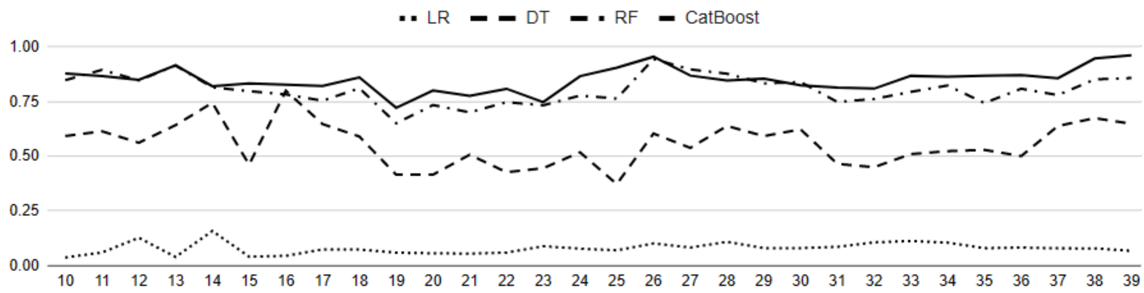

Figure 3. Classifier's AUC-PS behavior during the period of 30 quarters (10 to 39).

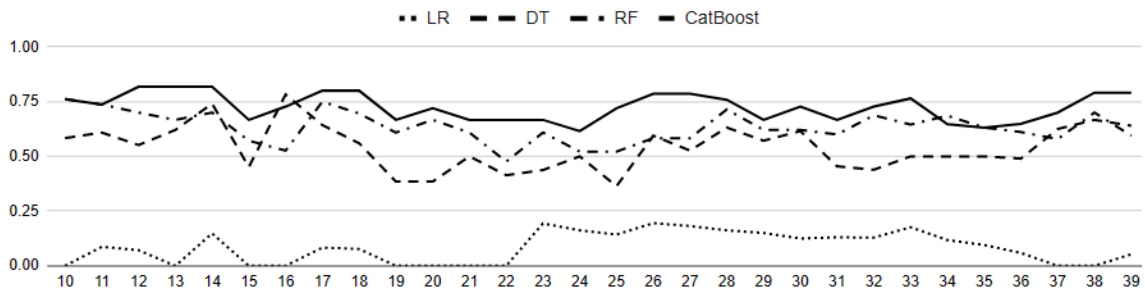

Figure 4. Classifier's F_1-Score behavior during the period of 30 quarters (10 to 39).

the other classifiers, except for the AUC-ROC metric where it showed similar behavior to DT.

The RF and CatBoost classifiers were markedly superior to LR and DT in terms of AUC-ROC and AUC-PS. Figure 2 shows that the worst values for RF and CatBoost were 0.9159 in the 19th quarter and 0.9611 in the 28th quarter, respectively, whereas the best values were 0.9980 in the 26th quarter for both classifiers. The best values obtained by LR and DT were 0.8291 in the 28th quarter and 0.8812 in the 14th quarter, respectively, which were still inferior to the worst values from RF and CatBoost and were less stable than these. Based on the AUC-PS time evolution curve (Figure 3), RF and CatBoost easily outperformed the others classifiers. Although RF and CatBoost classifiers exhibited similar behavior, CatBoost was slightly superior, with its lowest value of 0.7214 in the 19th quarter and the highest value of 0.9615 in the 39th quarter. In contrast, RF obtained 0.6505 in the same quarter and 0.9445 in the 26th quarter, respectively. Thus, the CatBoost had better performance than the other classifiers.

7. Conclusion

The present study presents a novel dataset of non-stationary data designed for FDP. The data is described by 84 economic-financial indicators and covers a period of 10-years organized into 40 quarters. This data were collected from the open data portal of CVM and organized considering the potential importance of the temporal dimension. To validate the dataset, we conducted experiments using a methodology that takes into account the time dependency of the data to deal with concept drift, outdated data, and class imbalance. Four different classifiers, namely LR, DT, RF, and CatBoost, were employed in this analysis, and their performances were measured using four metrics: AUC-ROC, AUC-PS, F_1-Score and G_{mean}.

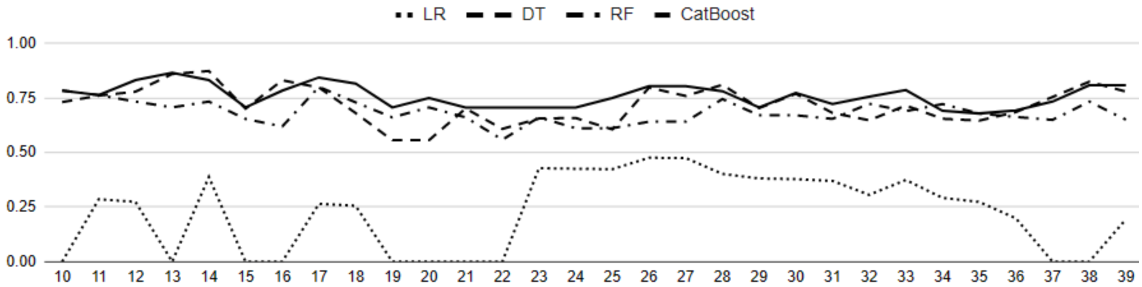

Figure 5. Classifier's G_{mean} behavior during the period of 30 quarters (10 to 39).

The results demonstrate that RF and CatBoost provide the most accurate AUC-ROC values (0.96691 and 0.98265) and the highest AUC-PS values (0.80441 and 0.85016), which indicates a good to excellent performance and well above random classifier [Saito and Rehmsmeier 2015]. When comparing these results with a similar study [Barboza et al. 2017], which focused on stationary data, this study achieved slightly better AUC-ROC rates for RF and the boosting method. Thus, the real-world dataset developed for this study can be utilized for FDP and has the potential to validate prediction models for the development of autonomous solutions.

Future studies should aim to identify occurrences of concept drift and determine its type in order to determine the best method for dealing with it while using fewer computational resources and achieving better response times. Due to the dataset's strong data imbalance characteristics, the use of data balancing techniques could improve the results, particularly for those with low AUC-PS rates (*i.e.* LR and DT). Additional data could be added to the dataset, which currently covers the period from 2011 to 2020, and a solution for automatic quarterly data collection from CVM could be implemented. For this purpose, the dataset used in this study is available on GitHub[15].

Acknowledgment

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

References

Agarwal, V. and Taffler, R. (2008). Comparing the performance of market-based and accounting-based bankruptcy prediction models. *Journal of Banking & Finance*, 32(8):1541–1551.

Alam, T. M., Shaukat, K., Mushtaq, M., Ali, Y., Khushi, M., Luo, S., and Wahab, A. (2020). Corporate Bankruptcy Prediction: An Approach Towards Better Corporate World. *The Computer Journal*, 64(11):1731–1746.

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

Altman, E. I. (2013). Predicting financial distress of companies: revisiting the z-score and zeta® models. In *Handbook of research methods and applications in empirical finance*, page 428–456. Edward Elgar Publishing.

[15] `https://github.com/rubensmchaves/ml-fdp`

Altman, E. I., Haldeman, R. G., and Narayanan, P. (1977). Zeta$^{\text{TM}}$ analysis a new model to identify bankruptcy risk of corporations. *Journal of banking & finance*, 1(1):29–54.

Altman, E. I., Marco, G., and Varetto, F. (1994). Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the italian experience). *Journal of Banking & Finance*, 18(3):505–529.

Barboza, F., Kimura, H., and Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, 83:405–417.

Barboza, F. L. d. M., Duarte, D. L., and Cunha, M. A. (2022). Anticipating corporate's distresses. *EXACTA Engenharia de Produção*, 20(2).

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, 4:71–111.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159.

Bragoli, D., Ferretti, C., Ganugi, P., Marseguerra, G., Mezzogori, D., and Zammori, F. (2022). Machine-learning models for bankruptcy prediction: do industrial variables matter? *Spatial Economic Analysis*, 17(2):156–177.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification And Regression Trees*. CRC Press, 1st edition edition.

Chen, Y., Guo, J., Huang, J., and Lin, B. (2022). A novel method for financial distress prediction based on sparse neural networks with $l_{1/2}$ regularization. *International Journal of Machine Learning and Cybernetics*, 13(7):2089–2103.

Cieslak, D. A. and Chawla, N. V. (2008). Learning decision trees for unbalanced data. In *Machine Learning and Knowledge Discovery in Databases*, pages 241–256.

Clement, C. (2020). Machine learning in bankruptcy prediction - a review. *Journal of Public Administration, Finance and Law*, 17:178–197.

Comissão de Valores Monetários (2022). Resolução CVM Nº 155, de 23 de Junho de 2022. *Diário Oficial da União*.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1—30.

Douglas, W. O. and Bates, G. E. (1933). The federal securities act of 1933. *Yale Law Journal*, 43(2):171.

Duarte, D. L. and Barboza, F. L. d. M. (2020). Forecasting financial distress with machine learning – a review. *Future Studies Research Journal: Trends and Strategies*, 12(3):528—-574.

Duarte, F. and Jones, C. (2017). Empirical network contagion for u.s. financial institutions. *FRB of NY Staff Report*, 1(826).

Eichengreen, B., Mody, A., Nedeljkovic, M., and Sarno, L. (2012). How the subprime crisis went global: Evidence from bank credit default swap spreads. *Journal of International Money and Finance*, 31(5):1299–1318.

Fernández, A., García, S., Galar, M., Prati, R., Krawczyk, B., and Herrera, F. (2018). *Learning from Imbalanced Data Sets*. Springer Cham.

Frydman, H., Altman, E. I., and Kao, D.-L. (1985). Introducing recursive partitioning for financial classification: the case of financial distress. *The journal of finance*, 40(1):269–291.

Gomes, H. M., Read, J., Bifet, A., Barddal, J. P., and Gama, J. (2019). Machine learning for streaming data: State of the art, challenges, and opportunities. *ACM SIGKDD Exploration Newsletter*, 21(2):6—-22.

Hanley, J. and Mcneil, B. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143:29–36.

Hui, X.-F. and Sun, J. (2006). An application of support vector machine to companies' financial distress prediction. In *Modeling Decisions for Artificial Intelligence*, pages 274–282. Springer Berlin Heidelberg.

Hyndman, R. J. and Athanasopoulos, G. (2021). *Forecasting: Principles and Practice*. OTexts.

Jabeur, S. B., Gharib, C., Mefteh-Wali, S., and Arfi, W. B. (2021). Catboost model and artificial intelligence techniques for corporate failure prediction. *Technological Forecasting and Social Change*, 166:120658.

Kim, H., Cho, H., and Ryu, D. (2022). Corporate bankruptcy prediction using machine learning methodologies with a focus on sequential data. *Computational Economics*, 59(3):1231–1249.

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

Kumbure, M. M., Lohrmann, C., Luukka, P., and Porras, J. (2022). Machine learning techniques and data for stock market forecasting: A literature review. *Expert Systems with Applications*, 197:116659.

Li, Z., Huang, W., Xiong, Y., Ren, S., and Zhu, T. (2020). Incremental learning imbalanced data streams with concept drift: The dynamic updated ensemble algorithm. *Knowledge-Based Systems*, 195:105694.

Liang, D., Lu, C.-C., Tsai, C.-F., and Shih, G.-A. (2016). Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European Journal of Operational Research*, 252(2):561–572.

Liang, D. and Tsai, C.-F. (2020). Taiwanese bankruptcy prediction. UCI Machine Learning Repository.

Lin, X., Zhang, Y., Wang, S., and Ji, G. (2013). A rule-based model for bankruptcy prediction based on an improved genetic ant colony algorithm. *Mathematical Problems in Engineering*, page 753251.

Lombardo, G., Pellegrino, M., Adosoglou, G., Cagnoni, S., Pardalos, P. M., and Poggi, A. (2022). Machine learning for bankruptcy prediction in the american stock market: Dataset and benchmarks. *Future Internet*, 14(8).

Martin, D. (1977). Early warning of bank failure: A logit regression approach. *Journal of Banking & Finance*, 1(3):249–276.

Martorano, L. (2021). Company bankruptcy prediction. Kaggle. Accessed: 2022-05-21.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18(1):109–131.

Pilch, B. (2021). An analysis of the effectiveness of bankruptcy prediction models – an industry approach. *Folia Oeconomica Stetinensia*, 21(2):76–96.

Ross, S. A., Westerfield, R., and Jaffe, J. (2012). *Corporate Finance*. Irwin/McGraw-Hill, 10th edition.

Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10:1–21.

Shen, F., Liu, Y., Wang, R., and Zhou, W. (2020). A dynamic financial distress forecast model with multiple forecast results under unbalanced data environment. *Knowledge-Based Systems*, 192:105365.

Shi, Y. and Li, X. (2019). A bibliometric study on intelligent techniques of bankruptcy prediction for corporate firms. *Heliyon*, 5(12):12.

Silva, T. C., da Silva Alexandre, M., and Tabak, B. M. (2017). Bank lending and systemic risk: A financial-real sector network approach with feedback. *Journal of Financial Stability*, 38:98–118.

Simon, C. J. (1989). The effect of the 1933 securities act on investor information and the performance of new issues. *The American Economic Review*, 79(3):295–318.

Succurro, M. (2017). Financial bankruptcy across european countries. *International Journal of Economics and Finance*, 9(7):132–146.

Sun, J., Li, H., Huang, Q.-H., and He, K.-Y. (2014). Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. *Knowledge-Based Systems*, 57:41–56.

Sun, J., yue Jia, M., and Li, H. (2011). Adaboost ensemble for financial distress prediction: An empirical comparison with data from chinese listed companies. *Expert Systems with Applications*, 38(8):9305–9312.

Sun, J., Zhou, M., Ai, W., and Li, H. (2019). Dynamic prediction of relative financial distress based on imbalanced data stream: from the view of one industry. *Risk Management*, 21(4):215–242.

Tang, Y., i, J., Zhu, Y., Gao, S., Tang, Z., and Todo, Y. (2019). A differential evolution-oriented pruning neural network model for bankruptcy prediction. *Complexity*, 2019(8682124).

Tomczak, S. (2016). Polish companies bankruptcy data. UCI Machine Learning Repository. DOI: `10.24432/C5F600`.

Wang, S., Minku, L. L., and Yao, X. (2018). A systematic study of online class imbalance learning with concept drift. *IEEE Transactions on Neural Networks and Learning Systems*, 29(10):4802–4821.

Zhang, H., Li, Z., Shahriar, H., Tao, L., Bhattacharya, P., and Qian, Y. (2019). Improving prediction accuracy for logistic regression on imbalanced datasets. In *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, volume 1, pages 918–919.

Zibanezhad, E., Foroghi, D., and Monadjemi, A. (2011). Applying decision tree to predict bankruptcy. In *2011 IEEE International Conference on Computer Science and Automation Engineering*, volume 4, pages 165–169.

Zieba, M., Tomczak, S. K., and Tomczak, J. M. (2016). Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert Systems with Applications*, 58:93–101.

Zou, Y., Gao, C., and Gao, H. (2022). Business failure prediction based on a cost-sensitive extreme gradient boosting machine. *IEEE Access*, 10:42623–42639.