Early warning of judicial recovery in Brazilian listed companies

Survival, machine learning and deep learning models on public filings, with distress signals read by language models

Rodrigo Hermont Ozon

2026-11-22

Abstract

We ask whether survival, machine learning and deep learning models anticipate judicial or out-of-court recovery of Brazilian listed non-financial companies better than the classical accounting scores, and whether distress signals read by language models from material-fact filings add information. The event base is built from public CVM records with point-in-time dating (103 events in 87 firms since 2011), the firm-quarter panel holds 33,037 observations of 1,007 firms, and eight models are compared on a single out-of-time test from 2020 onwards, with firm-clustered bootstrap intervals. Gradient-boosted and random-forest survival models reach an out-of-time AUC of about 0.85 for the 12-month PD, against 0.78 for the Altman score and 0.76 for the hazard logit; the neural networks do not beat the benchmarks, and no model is calibrated at every horizon. The pre-registered test finds no gain from the language-model signals (AUC 0.756 against 0.758); an exploratory rerun that keeps the rarest signal counts, which the registered preprocessing had erased, gives 0.783. A systematic review under PRISMA 2020 (4,019 records screened, 1,642 included) shows that survival models are a minority of the literature, that out-of-time validation is rarely stated, and that no study included combines survival models with language-model signals on Brazilian firms. The paper closes with an agentic credit memo whose every number is checked against the tools that produced it, and with a design for a corporate monitoring service aligned with the expected-loss rules of CMN Resolution 4,966/2021.

Resumo. Este artigo pergunta se modelos de sobrevivência, de aprendizado de máquina e de aprendizado profundo antecipam a recuperação judicial ou extrajudicial de companhias abertas brasileiras não financeiras melhor do que os escores contábeis clássicos, e se sinais de estresse lidos por modelos de linguagem nos fatos relevantes acrescentam informação. A base de eventos vem de registros públicos da CVM, datados pelo que era conhecido em cada data; oito modelos são comparados num único teste fora do tempo, a partir de 2020, com intervalos por bootstrap agrupado por empresa. Fora do tempo, os modelos de árvores chegam a uma AUC de cerca de 0,85 na PD de 12 meses, contra 0,78 do escore de Altman; nenhum modelo é bem calibrado em todos os prazos. O teste pré-registrado não encontra ganho nos sinais lidos pelos modelos de linguagem; uma rodada exploratória que mantém os sinais mais raros sugere um ganho que precisa de um novo período de teste para ser confirmado. Uma revisão sistemática (PRISMA 2020) mostra as lacunas que o estudo ocupa. Por fim, um parecer de crédito gerado por agentes, com trava numérica, e um desenho de serviço de monitoramento alinhado à Resolução CMN 4.966/2021.

1 Introduction

Judicial recovery filings by large Brazilian listed companies have become frequent enough to matter for every lender, supplier and bond investor in the country. Oi filed in June 2016 and again in January 2023; Americanas in January 2023, after an accounting inconsistency was disclosed; Light in May 2023. Each case was followed by losses for banks, bondholders and the suppliers who had sold on credit. The question a credit committee asks after each of them is the same: was there anything in the public record, a year earlier, that should have changed the decision?

The academic literature offers two traditions to answer it. Accounting scores, from Beaver (1966), Altman (1968) and Ohlson (1980) to the emerging-market version of Altman (2005), summarise the balance sheet in one number. Hazard models, introduced to the field by Shumway (2001) and extended by Chava and Jarrow (2004), Duffie et al. (2007) and CAMPBELL et al. (2008), treat default as an event in time and use every firm-year, not one snapshot per firm. Since then, machine learning (Barboza et al. 2017; Jones et al. 2016; Lessmann et al. 2015) and deep learning (Mai et al. 2019) have been compared with both, and text from filings has entered as a predictor (LOUGHRAN and MCDONALD 2011; Mai et al. 2019). Large language models now make it cheap to read every filing a company sends to its regulator (Huang et al. 2023; Kim et al. 2024; Lopez-Lira and Tang 2023).

Three things are missing for Brazil, and they are the subject of this paper. First, there is no public, point-in-time event base of judicial and out-of-court recoveries of listed firms; studies use the current registry status, which mixes dates and loses firms that left the market. Second, survival, tree-based and neural models have not been compared on the same Brazilian panel and the same out-of-time test. Third, no study we found measures, with a test fixed in advance, whether signals read by language models from material-fact filings improve on accounting and macroeconomic information.

We state three hypotheses. H1: survival and machine learning models discriminate better, out of time, than the Altman score and the discrete-time logit. H2: distress signals read by a language model from material facts improve the discrimination of the 12-month probability of default (PD). H3: survival models give better calibrated PD term structures at 12, 24 and 36 months, which is what lifetime expected loss requires. The primary test of H2 was fixed before the test period was opened (Section 5.5).

The paper is organised as follows. Section 2 summarises the regulatory frame. Section 3 reports the systematic review and the gaps it shows. Section 4 describes the data and the event base, Section 5 the method, and Section 6 the results. Section 7 presents the agentic credit memo and Section 8 the corporate solution. Section 9 lists the limitations and Section 10 concludes.

2 Regulatory context

CMN Resolution 4,966 of 25 November 2021 (Conselho Monetário Nacional 2021) moved Brazilian financial institutions to an expected-loss model for credit, with most of its provisions in force from 1 January 2025. Article 37 allocates each financial instrument to one of three stages: instruments without a significant increase in credit risk since origination (first stage), instruments whose credit risk has increased significantly (second stage), and instruments with a credit recovery problem (third stage). Under Article 47, the provision for the first stage is the expected loss over the next 12 months; for the second stage it is the expected loss over the whole expected term of the instrument. Article 40 asks the probability of default to consider the expected term and "reasonable and justifiable forecasts" of economic conditions.

Two requirements follow for a corporate credit model. It must produce a term structure of

default probabilities, not a single score, and it must detect a significant increase in credit risk early enough to move an exposure from the first to the second stage. Survival models produce the first by construction; the second is the early-warning problem this paper studies.

3 Systematic review

3.1 Protocol and search

The review follows PRISMA 2020 (Page et al. 2021). The protocol (`slr/protocol.md`) was written and committed before the search ran, and two dated amendments record the changes made while it ran. Three strands were searched for 2000 to 2026: (a) corporate default, bankruptcy or financial distress prediction with survival, machine learning or neural models; (b) text, natural language processing or large language models applied to credit risk or distress; (c) Brazilian evidence on bankruptcy, insolvency or judicial recovery. The protocol named OpenAlex as the source; when its anonymous quota was exhausted on the day of the search, the same strings were run on the Semantic Scholar bulk search (Amendment 1). Records were deduplicated by DOI and by normalised title.

The search returned 4,137 records; 118 duplicates were removed and 4,019 titles and abstracts were screened.

3.2 Screening and extraction

Titles and abstracts were screened against written criteria by two open-weight language models working independently (Gemma 4 26B and Gemma 4 31B, temperature zero). A record entered the review when either model included it (Amendment 2); this inclusive rule favours recall at the screening stage, where a false exclusion cannot be recovered later.

The two screeners agreed on 91.6% of the records (Cohen's kappa 0.82, (Cohen 1960)). 2,377 records were excluded and 1,642 included, of which 805 have an open-access full text.

Every included study was coded from its abstract: model families, validation design, country, unit of analysis, use of text and of language models, and number of events. A core of studies was then read in full. The core was drawn by stratum, so that each strand the gaps are about (survival models, deep learning, text and language models, Brazil, out-of-time validation) has up to ten studies read in depth, and was filled with the most cited studies per year since publication. Full texts were taken only from open-access copies; publisher sites that refuse automated requests were not worked around, and the next study in the stratum took the place of one that could not be retrieved.

Of 174 open-access studies tried, 101 gave a full text; the core holds 60 studies, 60 of them read and coded.

3.3 What the literature does and does not do

Table 1: Included studies by model family and period (a study can use several families).

Model family	2000-2009	2010-2019	2020-2026
Statistical (discriminant, logit, probit)	25	83	334
Trees and boosting	10	62	483
Support vector machines	10	39	117
Shallow neural networks	49	119	222
Deep learning	2	20	239
Survival models	6	31	59

Model family	2000-2009	2010-2019	2020-2026
Ensembles and hybrids	9	38	141
Text and language models	2	18	151
Other	12	22	68

Survival models appear in 96 of the 1,642 included studies, and only 9 combine them with deep learning. Text enters as a predictor in 204 studies and language models in 71; survival models and language models meet in 1 study. Of the 59 studies with Brazilian data, 1 uses a survival model, 0 deep learning and 0 a language model. Only 31 abstracts state an out-of-time test, and 1,482 do not say how the models were validated. The source gives no abstract for 618 of the included records, which were coded from the title alone, so every count here is a lower bound.

An abstract that does not describe the validation design does not prove the study lacked one, so the full-text core is the better estimate (Table 2). It is not reassuring. Of the 60 studies read in full, 32 still do not say how the test set was formed, 14 test on a later period (and the core over-samples them by design), 10 report any measure of calibration and 13 state the prediction horizon. The pattern supports the three gaps of the introduction: comparisons of model families rarely share one out-of-time test, survival models remain a small part of a field dominated by classification, calibration, which expected-loss provisioning depends on, is seldom checked, and the Brazilian literature has not yet used survival, deep learning or language models on listed firms.

Table 2: Full-text core: validation design and reporting.

Stratum	Studies	Out-of-time	Calibration reported	Uses text	Uses LLM
Brazil	10	1	0	0	0
Deep learning	10	0	2	6	3
Most cited (fill)	11	2	1	0	0
Out-of-time validation	9	6	5	1	1
Survival models	10	4	0	1	1
Text and language models	10	1	2	10	2
All	60	14	10	18	7

4 Data

4.1 Sources

All data are public. Financial statements come from the quarterly (ITR, 2011 to 2026) and annual (DFP, 2010 to 2026) filings in the CVM open-data portal (Comissão de Valores Mobiliários 2026), using consolidated accounts and individual accounts when no consolidated set exists. Each filing carries its delivery date to the regulator, which is the date the information becomes available to the model; a restated filing counts only from its own delivery date. The auditor's report attached to each filing is read for the going-concern paragraph. Event filings (IPE, 2003 to 2026) give the material facts and the category of documents filed by companies in judicial or

out-of-court recovery. Macroeconomic series come from the Central Bank's time-series system (Banco Central do Brasil 2026): the Selic rate, 12-month IPCA inflation, the IBC-Br activity index, unemployment, the corporate non-performing-loan ratio over 90 days and the average lending spread, each lagged by its publication delay.

4.2 Event base

An event is a filing for judicial recovery, a filing or court approval of out-of-court recovery, or a bankruptcy. Candidate events come from three sources: the registry status, the recovery category of event filings and material facts whose subject mentions recovery or bankruptcy. Material facts are adjudicated by a language model (Qwen3-235B, temperature zero) into filings by the company itself, ongoing proceedings, exits, or other matters such as a customer's or a subsidiary's recovery. An episode opens at the company's own filing and closes at its exit; a gap of more than a year, or an escalation from out-of-court to judicial recovery or to bankruptcy, opens a new episode. Liquidations are excluded. The base holds 129 events in 111 non-financial firms, 103 of them from 2011; 8 firms already in recovery on 1 January 2011 are left out.

Each draft event from 2008 on (130 events) is then checked by a second reader under a fixed rule. A language model (Gemma 4 31B, temperature zero) reads up to four of the company's filings nearest the event date and answers whether the company itself filed, and on what date. It is given the company's tax number, so that a filing made under a former name counts as the same company. An event the reader confirms, or dates within seven days, stands; one it dates elsewhere moves to that date, or goes if the company already has an event within seven days of the new date; one it attributes to another company (a parent, a subsidiary, an affiliate) goes; where it cannot tell, the first decision stands. Of the 130 events, 97 stand (18 of them because the second reader could not tell), 16 move and 17 go. The counts above are after this check.

4.3 Panel

The unit is the firm-quarter. For each calendar quarter, each firm carries the latest information delivered before the quarter ends; rows whose latest statement is more than 550 days old are dropped. The panel holds 33,037 firm-quarters of 1,007 firms from 2011-03-31 to 2026-09-30. Firms that left the market without an event are censored at their exit, so the sample has no survivorship bias. The training period ends in December 2018 (15,563 rows, 43 events); 2019 is an embargo year, so that no 12-month outcome window of the training rows overlaps the test; the test runs from 2020 (15,627 rows, 44 events).

5 Method

5.1 Predictors

Accounting variables follow the literature (Altman 1968; Ohlson 1980; Shumway 2001; CAMPBELL et al. 2008): working capital, retained earnings and EBIT over total assets, equity over total liabilities, leverage, current ratio, cash over assets, interest coverage, margin, revenue growth, size (assets deflated by IPCA), a negative-equity flag and a flag for a loss in the last four quarters. Flows are trailing twelve months. Disclosure behaviour adds the filing delay, the number of restatements in the last four quarters, the age of the latest information, the going-concern flag from a text rule on the auditor's report and an indicator of a previous recovery. The six macroeconomic series complete the base set.

5.2 Models

Eight models are fitted on the same rows and the same predictors: the Altman (2005) score with a logistic calibration; a discrete-time hazard logit (Shumway 2001) with an L2 penalty chosen by temporal cross-validation; a Cox model (Cox 1972) on landmark rows with a Breslow baseline (Breslow 1974); an XGBoost classifier of the 12-month PD and an XGBoost survival model with the Cox objective (Chen and Guestrin 2016); a random survival forest (Ishwaran et al. 2008); DeepSurv (Katzman et al. 2018); and DeepHit (Lee et al. 2018). Survival models are fitted on stacked landmarks: each firm-quarter starts a follow-up that ends at the event, at the firm's exit or at 36 months, which gives PD curves at 12, 24 and 36 months for every row. Hyper-parameters were set by a split inside the training years (fit to 2015, validate on 2017 and 2018); the test period was not used.

5.3 Evaluation

For the 12-month PD we report the area under the ROC curve (AUC), the area under the precision-recall curve, the Kolmogorov-Smirnov statistic, the Brier score and the calibration intercept and slope (Van Calster et al. 2019). For the term structure we report Harrell's (Harrell 1982) and Uno's (Uno et al. 2011) concordance, time-dependent AUC at 12, 24 and 35 months and the integrated Brier score (Graf et al. 1999). Intervals are 95% percentiles of a bootstrap that resamples firms, not rows, because rows of one firm are dependent. Differences in AUC are tested with DeLong's method (DeLong et al. 1988), with Holm's correction (Holm 1979) for the secondary comparisons.

5.4 Language-model signals

The subject line of each of the 30,990 material facts filed by the firms in the panel since 2010 is classified into eleven categories of distress: debt restructuring, covenant breach or waiver, missed payment, asset sale for liquidity, emergency capital, going-concern doubt, change of chief executive or chief financial officer, rating downgrade, the company's own recovery, a subsidiary's recovery, and sanction. The company's name, its trade names and its tax number are masked before the text reaches the model, so that the model's memory of what later happened to a known company cannot leak into the label. Announcements of the company's own recovery are kept for auditing the event base but never used as a predictor. A document enters the panel only from its delivery date; the features are counts of each category in the 4 and 12 months before the quarter. 7,181 documents carry at least one signal.

Reliability is measured two ways. A second model reads a random subsample of 3,000 documents and agreement is reported per category as Cohen's kappa (Cohen 1960), next to the number of flags from each reader, since on a rare category a few extra flags move kappa a long way. And 200 documents drawn at random are read in full, to measure what the subject line misses.

5.5 Pre-registered test of H2

The primary test of H2 was fixed before the test period was opened: the discrete-time hazard logit with and without the language-model signals, compared by the AUC of the 12-month PD on the out-of-time test, with a one-sided DeLong test at the 5% level. All other models, metrics and horizons are secondary and carry Holm's correction. The going-concern flag from the text rule stays in the baseline, so that the treated arm adds only what the language model reads.

The second reader's labels came in after the registration and before the test was opened. They led to one amendment, committed with them: a secondary version of the H2 test in which the treated arm keeps only the categories where the two readers reach a kappa of 0.60, the lower edge of what Landis and Koch (1977) call substantial agreement. The threshold was set after

the agreement was seen; the primary test is unchanged. A second amendment, also before the test was opened, replaced the review of the event list by hand with the second reader's rule of Section 4.2, so that the event base can be rebuilt by anyone from the cached answers.

6 Results

6.1 Discrimination and calibration (H1, H3)

On the 13,245 test firm-quarters whose 12-month window is observed (128 of them followed by a filing within 12 months), the tree ensembles rank firms best: XGBoost survival AUC 0.848 [0.799, 0.893]; XGBoost PD AUC 0.847 [0.797, 0.895]; Random survival forest AUC 0.846 [0.794, 0.894]; against 0.777 for the Altman score and 0.758 for the hazard logit (Table 3). XGBoost PD, XGBoost survival and Random survival forest beat both benchmarks after Holm's correction (Table 4); Cox beats the logit only; DeepSurv and DeepHit beat neither. H1 holds for the tree ensembles. The neural networks do not deliver the gain on a panel of this size, and the intervals overlap enough that the order among the three ensembles is not established.

No model is calibrated over the whole term structure (Table 5). At 12 months the observed rate is 1.25% and the mean PD ranges from 0.77% (XGBoost survival) to 2.87% (DeepHit); Hazard logit and DeepSurv come closest. At 36 months every model but DeepHit underestimates: the observed rate is 5.7% and the mean PDs run from 2.3% to 4.1%, while DeepHit overestimates by a wide margin. Calibration slopes are below one for every model except Cox, whose slopes above two mean its PDs are too compressed. Random survival forest has the lowest Brier score at 36 months. H3 holds only in part: the best long-horizon curve comes from a survival model, but survival models as a group are not better calibrated than the hazard logit, and every curve needs recalibration before it feeds a lifetime provision.

Table 3: Out-of-time test from 2020: discrimination of the 12-month PD (AUC with firm-clustered bootstrap interval, area under the precision-recall curve, KS, Brier score) and of the survival curve (Uno's C, integrated Brier score).

Model	AUC	95% CI	PR-AUC	KS	Brier	C (Uno)	IBS
Altman score	0.777	0.712 to 0.834	0.025	0.488	0.010	0.711	0.021
Hazard logit	0.758	0.674 to 0.834	0.034	0.479	0.010	0.711	0.022
Cox	0.804	0.740 to 0.861	0.043	0.523	0.009	0.736	0.021
XGBoost PD	0.847	0.797 to 0.895	0.053	0.576	0.010	0.795	0.022
XGBoost survival	0.848	0.799 to 0.893	0.071	0.556	0.009	0.807	0.021
Random survival forest	0.846	0.794 to 0.894	0.080	0.538	0.009	0.797	0.020
DeepSurv	0.798	0.728 to 0.862	0.045	0.502	0.010	0.733	0.020
DeepHit	0.763	0.689 to 0.838	0.057	0.466	0.011	0.612	0.057

Table 4: DeLong tests of the 12-month PD against the Altman score and the hazard logit (one-sided, model better), with Holm-adjusted p-values.

Model	Versus	AUC difference	p	p (Holm)
Cox	Altman score	0.027	0.031	0.154
Cox	Hazard logit	0.046	0.001	0.004
XGBoost PD	Altman score	0.070	0.000	0.000
XGBoost PD	Hazard logit	0.089	0.000	0.000
XGBoost survival	Altman score	0.071	0.000	0.000
XGBoost survival	Hazard logit	0.090	0.000	0.000
Random survival forest	Altman score	0.068	0.000	0.000
Random survival forest	Hazard logit	0.088	0.000	0.000
DeepSurv	Altman score	0.020	0.125	0.374
DeepSurv	Hazard logit	0.039	0.046	0.184
DeepHit	Altman score	-0.014	0.779	0.798
DeepHit	Hazard logit	0.005	0.399	0.798

Table 5: Calibration of the PD term structure on rows whose window is observed: mean PD and calibration slope at each horizon, and the Brier score at 36 months. Observed rates: 1.25% at 12 months, 3.06% at 24 and 5.71% at 36.

Model	PD 12m	Slope 12m	PD 24m	Slope 24m	PD 36m	Slope 36m	Brier 36m
Altman score	0.011	0.602	0.021	0.323	0.032	0.251	0.055
Hazard logit	0.012	0.737	0.025	0.633	0.037	0.517	0.055
Cox	0.012	2.416	0.024	2.196	0.037	1.913	0.053
XGBoost PD	0.008	0.589	0.016	0.504	0.023	0.447	0.055
XGBoost survival	0.008	0.694	0.018	0.592	0.032	0.527	0.053
Random survival forest	0.011	0.854	0.024	0.786	0.039	0.776	0.050
DeepSurv	0.013	0.827	0.026	0.699	0.041	0.564	0.052
DeepHit	0.029	0.660	0.050	0.595	0.289	0.021	0.245

6.2 Language-model signals (H2)

On the 3,000 subjects both models read, kappa ranges from 0.00 to 0.91 (Table 6). Raw agreement is at least 97.5% in every category, because most subjects carry no signal at all, so kappa is the informative figure. The first reader flags more often than the second in 9 of the 11 categories. Kappa falls below 0.60 in 3, each with many more flags from the first reader: going concern (11 against 1), subsidiary recovery (10 against 6) and sanction (111 against 46). The secondary H2 test leaves these categories out. In the full-text sample (197 documents read), the subject line finds 42 of the 67 documents that carry a signal in the full text (63%). The loss is uneven: asset sales (17 of 21) and executive changes (10 of 12) are named in the subject, while debt renegotiations mostly are not (3 of 14). Flags of the company's own recovery in the full text (14) come almost all (13) from firms already in recovery, whose filings restate that status. Since the

subject line keeps most of the signal, the panel uses subjects; the result for debt renegotiation is a limitation of H2 and the reason the monitor of Section 8 reads whole documents.

The pre-registered test does not support H2. On the 13,245 test rows (128 followed by a filing within 12 months), the hazard logit reaches an AUC of 0.758 without the signals and 0.756 with them (one-sided DeLong p = 0.58). With the seven reliable categories only, the AUC is 0.749 (p = 0.88). One flaw lies in the preprocessing, which was fixed with the rest of the code before the registration: every predictor is winsorised at the 1st and 99th percentiles of the training rows, and five signal counts (covenant waiver or breach, missed payment, going concern, rating downgrade and subsidiary recovery) are non-zero in fewer than 1% of those rows, so winsorising turned them into constant zeros and the registered tests ran without them. The tests were repeated with the signal counts left unclipped. That run was decided after the test period was opened and is exploratory: it gives an AUC of 0.783 with all signals (p = 0.01) and 0.777 with the reliable ones (p = 0.04). Across the other six models that take the signals, adding them changes the AUC by between -0.024 and 0.026, and none of the changes is significant after Holm's correction (Table 7).

Table 6: Agreement of the two language models on the material-fact subjects they both read: flags from each reader, flags from both, and Cohen's kappa.

Category	First reader	Second reader	Both	Kappa
Debt restructuring	65	59	47	0.75
Covenant waiver or breach	5	3	3	0.75
Missed payment	9	13	7	0.64
Asset sale for liquidity	194	163	153	0.85
Emergency capital	81	47	39	0.60
Going concern	11	1	0	0.00
Executive change	122	74	72	0.73
Rating downgrade	10	11	9	0.86
Own recovery	59	57	53	0.91
Subsidiary recovery	10	6	4	0.50
Sanction	111	46	41	0.51

Table 7: Each model with and without the language-model signals (registered preprocessing). The hazard logit's row is the pre-registered test; Holm's correction covers the seven secondary rows.

Model	Role	AUC base	AUC with signals	Difference	p	p (Holm)
Altman score	secondary	0.777	0.777	0.000	0.500	1.000
Hazard logit	primary	0.758	0.756	-0.002	0.583	
Cox	secondary	0.804	0.820	0.015	0.103	0.465
XGBoost PD	secondary	0.847	0.844	-0.003	0.734	1.000
XGBoost survival	secondary	0.848	0.857	0.009	0.030	0.210

Model	Role	AUC base	AUC with signals	Difference	p	p (Holm)
Random survival forest	secondary	0.846	0.854	0.008	0.050	0.297
DeepSurv	secondary	0.798	0.824	0.026	0.093	0.465
DeepHit	secondary	0.763	0.740	-0.024	0.910	1.000

6.3 Robustness

With the narrow event definition (judicial recovery and bankruptcy only, 92 test rows with an event within 12 months), Random survival forest has the highest AUC (0.857), against 0.803 for the Altman score and 0.784 for the hazard logit. Refitting every year with a moving origin (2021 to 2025), the mean AUC across years is highest for XGBoost survival (0.866) and Random survival forest (0.860); the Altman score averages 0.801 and the hazard logit 0.816. Yearly test sets hold few events, so single years move a lot (Table 8).

Table 8: AUC of the 12-month PD by test year, with an annual refit on a moving origin.

Model	2021	2022	2023	2024	2025
XGBoost survival	0.888	0.847	0.913	0.836	0.847
Random survival forest	0.862	0.816	0.884	0.852	0.887
XGBoost PD	0.886	0.864	0.905	0.813	0.830
Cox	0.915	0.774	0.828	0.848	0.772
Hazard logit	0.906	0.772	0.876	0.795	0.731
DeepSurv	0.961	0.789	0.767	0.812	0.729
DeepHit	0.975	0.806	0.859	0.736	0.658
Altman score	0.896	0.794	0.783	0.782	0.752

7 Agentic credit memo

The product layer turns the models into a credit memo for a committee. An orchestrator calls deterministic tools (the firm's panel row, its PD curve, the main drivers, the macroeconomic context and the material facts delivered before the memo date) and two language-model sub-agents: a reader, which summarises the recent filings, and a writer, which drafts the memo. A deterministic critic then checks every number and every date in the draft against the values the tools returned; a memo with a number or a year that no tool produced is blocked. Language models are known to state plausible facts that no source supports (Ji et al. 2023); the critic does not ask the model to behave, it removes the possibility.

The agent was run on four past cases (Americanas, Ambipar, Raízen, Casas Bahia), each as of twelve months before the filing, with the hazard logit fitted on the training years and only what was public on that date. The registry name, which now carries the company's later status, is cut back to the name alone, and other listed companies named in a memo are masked in the published copy. All four memos passed the critic. On the first run it blocked three, each for a value the memo had copied correctly: a negative rate read without its sign, and a crop year quoted from a filing subject; both cases are now tested. A year before the filing, the 12-month PD of Casas Bahia was 4.3%, about four times the average 12-month rate in the training years (1.1%); for Americanas, Ambipar and Raízen it was below that average (0.8%, 0.7%, 0.6%). Americanas filed days after disclosing an accounting inconsistency, so the statements a model

could read a year earlier did not show the problem. These runs illustrate the product and are not evidence for H2: the names are not masked in the memo, and the model may know how each story ended. The memos, in Portuguese, are in `results/memos/`.

8 Corporate solution

The design, called here a corporate credit monitor, follows from the regulatory frame and from what the experiment shows. It watches the CVM filings of the firms in a client's portfolio every day, recomputes a monthly PD curve per firm, raises an alert when a distress signal appears or the PD crosses a threshold set by the client's credit policy, and drafts a memo with the agentic layer for the credit committee. Every number in a memo traces back to a filing or a model output. It runs in the client's environment, and the decision stays with people.

The monitor maps directly onto Resolution 4,966: the 12-month PD feeds first-stage provisions, the 36-month curve the lifetime loss of the second stage, and the alerts are candidates for the significant-increase-in-credit-risk test that moves an exposure between them. A typical engagement starts with a diagnosis of the client's portfolio and data, runs a pilot of six to eight weeks on that portfolio, and continues with monthly monitoring.

9 Limitations

The number of events is small, so intervals are wide and the comparison rests on the intervals, not on the ranking. Market data (share prices, credit spreads) were left out because no public, stable source covers the whole panel. The event base depends on how firms describe their own filings; the list was checked by a second language model under a fixed rule rather than by hand, so some events may still carry a wrong date or belong to another company of the group, and a narrow definition (judicial recovery and bankruptcy only) is reported as a robustness check. No human labels were collected for the signals: the two models agree on most categories, but how close either comes to a human reading is not measured. Screening and extraction in the review were done by language models; their agreement is reported, but no human screened the full set, and studies without an open-access text enter only through their abstracts. The signals are read from subject lines; the full-text sample measures what that costs. The registered preprocessing erased the rarest signal counts, so the test of H2 that keeps them is exploratory and needs a later test period to confirm it. The memos on past cases name firms whose stories are public, and the model that writes them may know how those stories ended.

10 Conclusion

Public filings are enough to rank Brazilian listed companies by their risk of judicial or out-of-court recovery a year ahead, and tree ensembles do it better than the accounting score and the hazard logit the literature uses as benchmarks: an out-of-time AUC between 0.846 and 0.848, against 0.777 and 0.758. The advantage holds under the narrow event definition and in four of the five years of the moving-origin refit. The neural networks do not add to it on a panel of this size. The levels are another matter: every model misses the long-horizon rate, so a lender that uses these curves for lifetime provisions under Resolution 4,966 has to recalibrate them on its own portfolio.

Whether language models add to accounting data is not settled. The test registered in advance says no (AUC 0.756 with the signals, 0.758 without). The same test with the rarest signals kept (covenant breaches, missed payments, going-concern doubts, rating downgrades and subsidiaries' recoveries) gives 0.783, but that analysis came after the test was opened and needs a new test

period before it counts. The memos on past cases show what a monitor can and cannot do. A year before its filing, the hazard logit put Casas Bahia at 4.3%, about four times the average 12-month rate; it could not see Americanas, whose problem sat in accounts that no outside model could read.

The monitor of the previous section follows from these results: rank the portfolio with a tree ensemble, recalibrate the curves on the client's own defaults, read whole filings rather than subject lines, and let a deterministic critic check every number a language model writes.

Data and code availability

Code, derived data, the cached answers of every language-model call, tests and the review's records are in the public repository `rhozon/corporate-credit-early-warning`. Raw CVM and Central Bank files are downloaded by the code. Language models were used as instruments of the method, as described in Sections 3 to 5 and 7; the models, versions and settings are recorded in the repository.

References

Altman, Edward I. 1968. "FINANCIAL RATIOS, DISCRIMINANT ANALYSIS AND THE PREDICTION OF CORPORATE BANKRUPTCY." *The Journal of Finance* 23 (4): 589–609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x.

Altman, Edward I. 2005. "An Emerging Market Credit Scoring System for Corporate Bonds." *Emerging Markets Review* 6 (4): 311–23. https://doi.org/10.1016/j.ememar.2005.09.007.

Banco Central do Brasil. 2026. *Sistema Gerenciador de Séries Temporais (SGS)*. https://www3.bcb.gov.br/sgspub/.

Barboza, Flavio, Herbert Kimura, and Edward Altman. 2017. "Machine Learning Models and Bankruptcy Prediction." *Expert Systems with Applications* 83: 405–17. https://doi.org/10.1016/j.eswa.2017.04.006.

Beaver, William H. 1966. "Financial Ratios as Predictors of Failure." *Journal of Accounting Research* 4: 71. https://doi.org/10.2307/2490171.

Breslow, N. 1974. "Covariance Analysis of Censored Survival Data." *Biometrics* 30 (1): 89. https://doi.org/10.2307/2529620.

CAMPBELL, JOHN Y., JENS HILSCHER, and JAN SZILAGYI. 2008. "In Search of Distress Risk." *The Journal of Finance* 63 (6): 2899–939. https://doi.org/10.1111/j.1540-6261.2008.01416.x.

Chava, Sudheer, and Robert A. Jarrow. 2004. "Bankruptcy Prediction with Industry Effects." *Review of Finance* 8 (4): 537–69. https://doi.org/10.1093/rof/8.4.537.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–94. https://doi.org/10.1145/2939672.2939785.

Cohen, Jacob. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and Psy-*

chological Measurement 20 (1): 37–46. https://doi.org/10.1177/001316446002000104.

Comissão de Valores Mobiliários. 2026. *Portal de Dados Abertos Da CVM: Cadastro de Companhias Abertas, ITR, DFP e IPE*. https://dados.cvm.gov.br/.

Conselho Monetário Nacional. 2021. *Resolução CMN n.º 4.966, de 25 de Novembro de 2021: Dispõe Sobre Os Conceitos e Os Critérios Contábeis Aplicáveis a Instrumentos Financeiros*. Banco Central do Brasil, normative search. https://www.bcb.gov.br/estabilidadefinanceira/exibenormativo?tipo=Resolu%C3%A7%C3%A3o%20CMN&numero=4966.

Cox, D. R. 1972. "Regression Models and Life-Tables." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 34 (2): 187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x.

DeLong, Elizabeth R., David M. DeLong, and Daniel L. Clarke-Pearson. 1988. "Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach." *Biometrics* 44 (3): 837. https://doi.org/10.2307/2531595.

Duffie, Darrell, Leandro Saita, and Ke Wang. 2007. "Multi-Period Corporate Default Prediction with Stochastic Covariates." *Journal of Financial Economics* 83 (3): 635–65. https://doi.org/10.1016/j.jfineco.2005.10.011.

Graf, Erika, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. 1999. "Assessment and Comparison of Prognostic Classification Schemes for Survival Data." *Statistics in Medicine* 18 (17-18): 2529–45. https://doi.org/10.1002/(sici)1097-0258(19990915/30)18:17/18%3C2529::aid-sim274%3E3.0.co;2-5.

Harrell, Frank E. 1982. "Evaluating the Yield of Medical Tests." *JAMA: The Journal of the American Medical Association* 247 (18): 2543. https://doi.org/10.1001/jama.1982.03320430047030.

Holm, Sture. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics* 6 (2): 65–70.

Huang, Allen H., Hui Wang, and Yi Yang. 2023. "<Scp>FinBERT</Scp> : A Large Language Model for Extracting Information from Financial Text*." *Contemporary Accounting Research* 40 (2): 806–41. https://doi.org/10.1111/1911-3846.12832.

Ishwaran, Hemant, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. 2008. "Random Survival Forests." *The Annals of Applied Statistics* 2 (3): 841–60. https://doi.org/10.1214/08-aoas169.

Ji, Ziwei, Nayeon Lee, Rita Frieske, et al. 2023. "Survey of Hallucination in Natural Language Generation." *ACM Computing Surveys* 55 (12): 1–38. https://doi.org/10.1145/3571730.

Jones, Stewart, David Johnstone, and Roy Wilson. 2016. "Predicting Corporate Bankruptcy: An Evaluation of Alternative Statistical Frameworks." *Journal of Business Finance &Amp; Accounting* 44 (1-2): 3–34. https://doi.org/10.1111/jbfa.12218.

Katzman, Jared L., Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. 2018. "DeepSurv: Personalized Treatment Recommender System Using a

Cox Proportional Hazards Deep Neural Network." *BMC Medical Research Methodology* 18 (1). https://doi.org/10.1186/s12874-018-0482-1.

Kim, Alex G., Maximilian Muhn, and Valeri V. Nikolaev. 2024. *Financial Statement Analysis with Large Language Models*. https://doi.org/10.2139/ssrn.4835311.

Landis, J. Richard, and Gary G. Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics* 33 (1): 159. https://doi.org/10.2307/2529310.

Lee, Changhee, William Zame, Jinsung Yoon, and Mihaela Van der Schaar. 2018. "DeepHit: A Deep Learning Approach to Survival Analysis with Competing Risks." *Proceedings of the AAAI Conference on Artificial Intelligence* 32 (1). https://doi.org/10.1609/aaai.v32i1.11842.

Lessmann, Stefan, Bart Baesens, Hsin-Vonn Seow, and Lyn C. Thomas. 2015. "Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: An Update of Research." *European Journal of Operational Research* 247 (1): 124–36. https://doi.org/10.1016/j.ejor.2015.05.030.

Lopez-Lira, Alejandro, and Yuehua Tang. 2023. *Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models*. https://doi.org/10.2139/ssrn.4412788.

LOUGHRAN, TIM, and BILL MCDONALD. 2011. "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-ks." *The Journal of Finance* 66 (1): 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x.

Mai, Feng, Shaonan Tian, Chihoon Lee, and Ling Ma. 2019. "Deep Learning Models for Bankruptcy Prediction Using Textual Disclosures." *European Journal of Operational Research* 274 (2): 743–58. https://doi.org/10.1016/j.ejor.2018.10.024.

Ohlson, James A. 1980. "Financial Ratios and the Probabilistic Prediction of Bankruptcy." *Journal of Accounting Research* 18 (1): 109. https://doi.org/10.2307/2490395.

Page, Matthew J, Joanne E McKenzie, Patrick M Bossuyt, et al. 2021. "The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews." *BMJ*, n71. https://doi.org/10.1136/bmj.n71.

Shumway, Tyler. 2001. "Forecasting Bankruptcy More Accurately: A Simple Hazard Model." *The Journal of Business* 74 (1): 101–24. https://doi.org/10.1086/209665.

Uno, Hajime, Tianxi Cai, Michael J. Pencina, Ralph B. D'Agostino, and L. J. Wei. 2011. "On the c-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data." *Statistics in Medicine* 30 (10): 1105–17. https://doi.org/10.1002/sim.4154.

Van Calster, Ben and, David J. McLernon, Maarten van Smeden, Laure Wynants, and Ewout W. Steyerberg. 2019. "Calibration: The Achilles Heel of Predictive Analytics." *BMC Medicine* 17 (1). https://doi.org/10.1186/s12916-019-1466-7.