

MACHINE LEARNING EARLY WARNING FOR FINANCIAL DISTRESS IN HEALTH PLAN OPERATORS

Alerta antecipado por aprendizado de máquina para dificuldades financeiras em operadoras de planos de saúde

AUTORES

Guilherme Coelho¹

Clarimar José Coelho²

1 Universidade Estadual de Campinas

2 Pontifícia Universidade Católica de Goiás

CONTATO

Guilherme Coelho

1 guilcoel@unicamp.br

Obs: este artigo reflete exclusivamente a opinião do(s) autor(es), não representando a posição do IESS.

ABSTRACT

This study develops the first machine-learning-based early-warning system for financial distress among Brazilian health plan operators. Using 24,440 operator-quarter observations from public regulatory data (2018–2025), we train and temporally validate LASSO logistic regression, random forest, and XGBoost to predict distress two to four quarters ahead. Random forest achieved the highest discrimination (AUC = 0.847), but LASSO exhibited the smallest generalization gap (0.014), revealing a tension between accuracy and deployment reliability that carries direct implications for regulatory design. The extended combined ratio was the only consensus predictor across all methods, and temporal features added predictive value beyond static ratio levels. A retrospective case study of Brazil's largest operator demonstrated early-detection capacity but also revealed model habituation under prolonged stress. The study extends early warning theory from banking to health insurance and demonstrates that open regulatory data can support proactive, transparent surveillance where operator failure directly affects healthcare access.

Keywords: Early Warning System, Financial Distress, Health Insurance Regulation, Machine Learning, Supplementary Health.

RESUMO

Este estudo desenvolve o primeiro sistema de alerta antecipado baseado em aprendizado de máquina para identificar dificuldades financeiras de operadoras de planos de saúde no Brasil. Com 24.440 observações operadoras-trimestre de dados regulatórios públicos (2018–2025), foram validados, com abordagem temporal, os modelos LASSO, random forest e XGBoost para prever com dois a quatro trimestres de antecedência. O random forest obteve maior discriminação (AUC = 0,847), mas o LASSO apresentou menor lacuna de generalização (0,014), o que revela tensão entre acurácia e confiabilidade regulatória. O índice combinado ampliado foi o único preditor consensual, e as variáveis temporais agregaram valor preditivo. Estudo de caso na maior operadora demonstrou detecção precoce, mas também habituação do modelo sob estresse prolongado. O trabalho estende a teoria do alerta antecipado à saúde suplementar e demonstra que dados públicos sustentam a vigilância proativa.

Palavras-chave: Sistema de Alerta Antecipado, Dificuldade Financeira, Regulação de Planos de Saúde, Aprendizado de Máquina, Saúde Suplementar.

INTRODUÇÃO

Brazil's supplementary health system covers approximately 51 million beneficiaries through a regulated market of private health insurance operators (Agência Nacional de Saúde Suplementar [ANS], 2025). The sector has experienced intense consolidation—from over 2,000 active operators in 2000 to fewer than 700 with enrolled beneficiaries by 2025—alongside persistent financial fragility (ANS, 2025; Ocké-Reis, 2012). Recent cases of large operators facing acute financial deterioration have crystallized these vulnerabilities, underscoring the need for forward-looking regulatory tools capable of anticipating distress before they materialize.

The ANS maintains a comprehensive prudential monitoring framework, including quarterly financial reporting through the disclosure of economic and financial information system (DIOPS) data, prudential classification scores, and the IDSS (ANS, 2024; ANS, 2022). However, the current regulatory approach is largely reactive: special regimes (fiscal direction or technical direction) are imposed after distress has materialized, and prudential classifications reflect lagging indicators. This regulatory gap motivates the development of forward-looking predictive tools.

Early warning systems (EWS) for financial institutions have a long tradition in banking regulation, exemplified by the CAMELS framework (Cole & Gunther, 1998; Sahajwala & Van den Bergh, 2000). These models combine financial ratios, operational metrics, and temporal dynamics to identify institutions at risk of failure before formal regulatory intervention (Altman, 1968; Ohlson, 1980). More recently, machine learning methods have substantially improved predictive accuracy for corporate distress, with ensemble models such as random forests achieving area under the curve (AUC) values above 0.85 in cross-country benchmarks (Barboza et al., 2017). The insurance sector has adapted these approaches, particularly in property-casualty and life insurance markets (Cummins et al., 1999; Eling & Pankoke, 2016).

A small but growing body of research has examined the financial performance of Brazilian health operators using accounting data and traditional econometric methods (Silva & Loebel, 2016; Bragança et al., 2019; Reis et al., 2021; Souza & Silva, 2020). However, no study has employed a forward-looking machine learning framework with temporal validation to predict financial distress before it occurs. This gap is consequential: the tools available to ANS for risk identification remain fundamentally backward-looking, leaving regulators and market participants without early signals of impending deterioration.

The Brazilian supplementary health system also presents distinctive structural features that render direct transposition of banking or insurance early warning systems (EWS) inadequate. Health-related litigation grew 112% between 2020 and 2024, reaching approximately 299,000 new lawsuits in 2024 alone, with accumulated costs of R\$ 17.1 billion over five years (Instituto de Estudos de Saúde Suplementar [IESS], 2025). The financial burden of this litigation—focused on coverage denials and high-cost medications—disproportionately affects smaller operators and directly contributes to deterioration in combined ratios (Trettel et al., 2018; Wang et al., 2024). In parallel, medical cost inflation has consistently exceeded general consumer price inflation (IESS, 2024), and demographic pressures from an aging beneficiary pool further strain the system. These features create a distinctive risk profile that requires sector-specific predictive models.

This study addresses three questions. First, can open regulatory data from ANS predict financial distress in health operators two to four quarters ahead? Second, which financial indicators carry the strongest predictive signal? Third, would the proposed model have flagged financial deterioration in a large national operator before distress became evident? Drawing on the EWS banking literature, we expect ensemble methods to outperform linear models due to the nonlinear nature of financial distress processes (Barboza et al., 2017), and anticipate that combined ratio-related indicators will dominate feature importance, consistent with actuarial theory on insurer solvency (Cummins et al., 1999). We contribute to the literature by (a) constructing the first machine learning-based EWS specifically designed for health insurance operators, (b) demonstrating that exclusively public data achieves meaningful predictive performance using an ensemble of least absolute shrinkage and selection operator (LASSO), random forest, and gradient boosting models, and (c) providing interpretable results through LASSO feature selection (Tibshirani, 1996) and Shapley Additive Explanations (SHAP) analysis (Lundberg & Lee, 2017) that can inform both regulatory policy and managerial decision-making.

THEORETICAL BACKGROUND

Financial Distress and Early Warning Systems

The prediction of financial distress has evolved through three foundational contributions: Beaver's (1966) demonstration that cash flow to total debt could discriminate failing firms up to five years ahead; Altman's (1968) multivariate Z-score; and Ohlson's (1980) logistic regression approach, which produced probability estimates

rather than binary classifications. A critical limitation of these models is their reliance on cross-sectional snapshots. Shumway (2001) demonstrated that discrete-time hazard models with time-varying covariates produced more accurate forecasts, showing that the temporal dynamics of financial indicators may be as informative as their levels—an insight with direct implications for EWS design in any regulated sector.

In banking regulation, EWS emerged to identify vulnerable institutions before crises materialize, with the CAMELS framework becoming the standard off-site monitoring approach (Sahajwala & Van den Bergh, 2000). Cole and Gunther (1998) demonstrated that statistical models could match or outperform expert-based CAMELS ratings. In insurance, Cummins et al. (1999) developed solvency prediction models using risk-based capital ratios, while Eling and Pankoke (2016) noted that traditional actuarial indicators often fail to capture interconnected vulnerabilities. Despite this tradition, no comparable EWS has been proposed for health insurance operators—a gap that is particularly consequential where operator failure directly affects healthcare access.

Machine Learning in Financial Prediction

Machine learning has challenged traditional econometric approaches to financial prediction. Barboza et al. (2017) found that ensemble methods achieved AUC values above 0.87, substantially outperforming logistic regression across a large U.S. sample. Lessmann et al. (2015) confirmed this pattern across 41 classifiers, attributing the advantage to tree-based methods' ability to capture nonlinear relationships without a priori specification. For regulatory applications, however, predictive accuracy alone is insufficient: supervisors must justify interventions with transparent evidence, creating a tension between performance and interpretability.

This study employs three algorithms spanning the interpretability–performance spectrum. LASSO logistic regression (Tibshirani, 1996) performs simultaneous estimation and variable selection through L1 penalization, identifying the most parsimonious predictor subset while producing interpretable coefficients. Random forests (Breiman, 2001) aggregate predictions from decorrelated decision trees, achieving low variance through ensemble averaging. XGBoost (Chen & Guestrin, 2016) implements regularized gradient-boosted trees. To address the interpretability gap of ensemble methods, we employ SHAP values (Lundberg & Lee, 2017), which assign each feature a consistent marginal contribution grounded in cooperative game theory.

The Brazilian Supplementary Health Sector

Brazil's supplementary health system operates through private operators covering approximately 51 million beneficiaries under ANS regulatory oversight (ANS, 2025). Operators submit quarterly financial statements through DIOPS, constituting a rich but underexploited panel dataset. The ANS monitors operators via composite indicators such as the IDSS, which aggregates financial, operational, and quality dimensions. However, the IDSS was not designed as a predictive instrument: it assigns arbitrary weights, does not model temporal dynamics, and produces overall assessments rather than probability estimates. When deterioration reaches critical thresholds, special regulatory regimes may be imposed under RN nº 510 (ANS, 2022), but this architecture operates primarily in reactive mode (Ocké-Reis, 2012).

A growing body of literature has examined the financial performance of Brazilian health operators. Silva and Loebel (2016) identified persistent performance differences among operator modalities but did not develop predictive models. Bragança et al. (2019) used logistic regression with 26 financial variables to predict liquidation, finding leverage and provisioning as the strongest predictors, but their model was not validated out-of-sample. Reis et al. (2021) extended this line using factor analysis followed by logistic regression with resampling, finding that liquidity, leverage, and financial slack predicted special regimes up to two years ahead—the closest existing approximation to a predictive EWS, though limited by reliance on a single linear method without temporal feature engineering.

The absence of forward-looking models is consequential. First, existing regulatory tools summarize past performance rather than estimating future risk trajectories; by the time operators enter special regimes, beneficiary harm has already occurred. Second, the sector's consolidation has increased systemic interdependence: large operator failures cascade through provider networks, disrupting healthcare access for millions of beneficiaries (ANS, 2024). A machine learning-based EWS, trained on publicly available DIOPS data, could provide the anticipatory capacity that existing tools lack, enabling a shift from reactive intervention to proactive risk management while maintaining regulatory accountability. Drawing on this review, we derive three testable propositions:

P1 (Model performance): Given the documented superiority of ensemble methods over linear classifiers in financial distress prediction (Barboza et al., 2017; Lessmann et al., 2015), we expect random forest and XGBoost to outperform LASSO logistic regression in predicting health operator distress, as measured by AUC-ROC.

P2 (Feature dominance): Based on actuarial theory linking insurer solvency to claims-paying capacity and prior evidence from the Brazilian sector (Bragança et al., 2019; Reis et al., 2021), we expect combined ratio-related indicators and liquidity measures to dominate feature importance rankings across all models.

P3 (Temporal dynamics): Given that financial deterioration is a process rather than a state (Shumway, 2001), we expect that trend-based features—such as slopes and volatility of key ratios over rolling windows—will add predictive value beyond static ratio levels.

METHODS

Data and Sample

We use quarterly financial statements submitted by health plan operators to ANS through the DIOPS, a standardized electronic reporting system mandated by regulatory resolution (*Resolução Normativa*) nº 510 (ANS, 2022). The DIOPS database is publicly available through the ANS data repository and contains balance sheet and income statement data for all active operators, updated quarterly. Our sample covers 30 quarters from the first quarter of 2018 (2018Q1) through the third quarter of 2025 (2025Q3), encompassing 679-879 medical-hospital operators per quarter (reflecting market entries and exits over the period) and yielding a panel of 24,440 operator-quarter observations. We restrict the sample to medical-hospital operators (*médico-hospitales*), which constitute the majority of the market and share a comparable regulatory structure. Operators with fewer than four consecutive quarters of data are excluded, as temporal feature construction requires a minimum observation window. Data from the beneficiaries' information system (*Sistema de Informações de Beneficiários*, SIB) were tested in a preliminary ablation analysis. However, they did not improve model performance and were excluded from the final specification.

Target Variable Definition

We define financial distress as a multicomponent binary variable (Y) that captures regulatory outcomes signaling severe operational deterioration. Specifically, $Y = 1$ if any of the following events occurs within a forward-looking window of $K = 4$ quarters: (a) cancellation of the operator's registration by ANS, (b) entry into a special regulatory regime—fiscal direction (*direção fiscal*) or technical direction (*direção técnica*)—as established by Law nº 9.656 (Brasil, 1998) and regulated by RN nº 316 (ANS, 2012), or (c) classification in prudential categories S3 or S4 under the risk-based framework

established by RN nº 443 (ANS, 2019). This composite definition is preferable to single-event targets for three reasons. First, these events represent stages of a single deterioration continuum: in ANS regulatory practice, S3/S4 classification frequently precedes fiscal direction, which may culminate in cancellation (Reis et al., 2021). Second, the low individual prevalence of each event type (with cancellations alone numbering fewer than 50 in the training set) would make separate modeling infeasible with 77 candidate features. Third, composite outcomes are standard in the banking EWS literature, where the CAMELS framework encompasses rating downgrades through liquidation (Cole & Gunther, 1998). The four-quarter horizon balances the utility of early detection against signal strength. Operator-quarter observations occurring after the event are excluded. In the training set, the prevalence of $Y = 1$ is approximately 1.7%, reflecting the rarity of severe distress events.

Feature Engineering

We construct 77 candidate features from DIOPS raw variables, organized into five substantive categories: (a) profitability indicators, including loss ratio (*sinistralidade*), combined ratio, extended combined ratio, and operating margin; (b) liquidity, represented by the current ratio; (c) leverage, measured as total liabilities relative to total assets; (d) expense structure, comprising administrative expense ratio and the ratio of financial expenses to financial revenue; and (e) equity dynamics, including equity-to-revenue ratio and absolute and percentage changes in equity. Beyond these contemporaneous ratios, we engineer temporal features that capture deterioration dynamics: quarter-over-quarter changes (d_1q) and four-quarter rolling statistics (minimum, maximum, standard deviation, slope, range, and percentage change). These temporal features are motivated by the theoretical insight that the trajectory of financial indicators carries predictive information beyond their current levels (Shumway, 2001), consistent with Proposition P3 advanced in the Theoretical Background. All features are derived exclusively from publicly available DIOPS data, ensuring full reproducibility and practical applicability for regulators and market participants (a complete list of features with descriptive statistics is available in the code repository).

Continuous features are winsorized at the 1st and 99th percentiles to limit the influence of extreme outliers. Missing values—primarily arising from rolling window calculations in the initial quarters of each operator's series (affecting less than 5% of observations)—are imputed as zero, reflecting a conservative assumption that the absence

of temporal variation signals stability. We note that this choice affects only the warm-up period for rolling features, and that a formal sensitivity analysis comparing three imputation strategies—zero imputation, complete exclusion of operators with fewer than eight quarters, and quarter-median imputation—yielded substantively identical results (see Results).

Models

We estimate three classification models that span a gradient of complexity and interpretability, all trained on the same data with class-imbalance correction (target weight ratio $w_{pos} = N_0/N_1 \approx 56.8$). The correction is implemented as observation weights in LASSO, as case weights in the random forest, and as `scale_pos_weight` in XGBoost, each following the respective package's native approach to cost-sensitive learning.

LASSO Logistic Regression

The L1-regularized logistic regression (Tibshirani, 1996) serves as the interpretable baseline. The regularization parameter λ is selected via 10-fold cross-validation on the training set using the one-standard-error rule, which favors a more parsimonious model over the minimum-deviance solution. This procedure retains 11 of the 77 candidate features, yielding coefficients directly interpretable as log-odds ratios. The LASSO addresses both variable selection and multicollinearity among the highly correlated financial ratios.

Random Forest

The random forest (Breiman, 2001) is configured with 1,000 trees, $mtry = \sqrt{p}$ (where $p = 77$ is the number of features), a minimum node size of 20, and a maximum tree depth of 8. These conservative hyperparameters are chosen to mitigate overfitting, given the small number of positive cases. Predicted probabilities are obtained from the proportion of trees voting for the positive class. Feature importance is quantified through impurity-based reduction (mean decrease in Gini index).

XGBoost

The gradient boosting model (Chen & Guestrin, 2016) is trained with a learning rate (η) of 0.05, maximum tree depth of 4, minimum child weight of 20, row subsampling at 80%, column subsampling at 70%, and L1 and L2 regularization terms of 0.1 and 1.0,

respectively. The optimal number of boosting rounds (382) is determined by 5-fold cross-validation with early stopping on log-loss. SHAP values (Lundberg & Lee, 2017) are computed natively within the XGBoost model to provide both global feature-importance rankings and local, observation-level explanations of predicted risk.

Validation Strategy

We employ a strict temporal split to prevent data leakage: all observations up to and including 2023Q3 constitute the training set, and all observations from 2024Q1 onward form the held-out test set. This design reflects the real-world deployment scenario in which models trained on historical data must predict future distress events without access to contemporaneous or future information. Model discrimination is evaluated using AUC-ROC with 95% confidence intervals estimated by the nonparametric method of DeLong et al. (1988). Pairwise comparisons between models use the same DeLong test. Additional metrics include sensitivity, specificity, precision, F1 score, and the Brier score, which captures calibration quality. Classification thresholds are selected by maximizing Youden's J statistic ($\text{sensitivity} + \text{specificity} - 1$) on the training set, rather than using the default 0.5 cutoff, which would be inappropriate given the low prevalence of positive cases. To assess generalization, we report the overfitting gap ($\text{AUC}_{\text{train}} - \text{AUC}_{\text{test}}$), where a larger gap indicates greater overfitting to the training data. Model robustness is further evaluated through an expanding-window cross-validation procedure following the rolling-origin framework recommended by Tashman (2000): the initial training window comprises the first six quarters (2018Q1–2019Q2), and each subsequent fold expands the training set by one quarter while testing on the next available quarter, yielding 16 folds (test quarters from 2019Q3 through 2023Q3). This procedure yields fold-level AUC estimates that capture the model's stability across different economic conditions, including the COVID-19 pandemic period. Comparative model performance is summarized in Table 1 and visualized in Figure 1.

Interpretability and Case Study

To assess the practical value of the EWS, we conduct a retrospective case study on a large national group medicine operator (*medicina de grupo*), Brazil's largest health plan operator by number of beneficiaries at the time of analysis. Using an expanding-window backtest from 2020Q1 through 2025Q3, we retrain each model at every quarter and score only the selected operator, producing a time series of predicted distress

probabilities. This design allows us to evaluate whether the models could have provided early warning signals of financial deterioration in a systemically important operator (Figure 5). Feature importance is compared across all three models (Table 4) to identify consensus predictors: features that rank consistently high across modeling approaches are more likely to reflect genuine risk signals rather than method-specific artifacts. SHAP values for the XGBoost model are presented in Figure 2, and LASSO coefficient paths in Figure 3.

Software, Reproducibility, and Ethical Considerations

All analyses were conducted in R (version 4.5.2) using the `glmnet` (LASSO), `ranger` (random forest), `xgboost`, and `pROC` (DeLong test) packages. A fixed random seed (`set.seed = 42`) was used in all stochastic procedures to ensure exact reproducibility. The complete analytical pipeline—from raw data download to final figures—comprises 10 scripts and is deposited at [repository URL blinded for review], along with the processed datasets. Raw DIOPS data are publicly available at the ANS data repository (<ftp.ans.gov.br>). This study uses exclusively publicly available, de-identified administrative data from a government regulatory agency and does not involve human subjects research; it is therefore exempt from institutional review board approval under Brazilian regulation (Conselho Nacional de Saúde [CNS], 2016).

RESULTS

Descriptive Statistics

The final analytical sample comprised 24,440 operator-quarter observations spanning 23 training quarters (2018Q1–2023Q3) and 7 test quarters (2024Q1–2025Q3). In the training set ($n = 17,812$), 312 observations (1.75%) were classified as distressed ($Y = 1$). In the test set ($n = 6,628$), 29 observations (0.44%) were positive, reflecting both the rarity of severe distress events and temporal variation in regulatory actions. The limited number of distress events in the test set ($n = 29$) introduces statistical uncertainty into performance estimates, particularly for threshold-dependent metrics such as precision and sensitivity. This limitation stems from the rarity of severe regulatory interventions and underscores the importance of confidence intervals and complementary validation approaches when interpreting model performance. Descriptive statistics for all 77 candidate features, comparing training and test sets on key financial indicators, are available in the code repository.

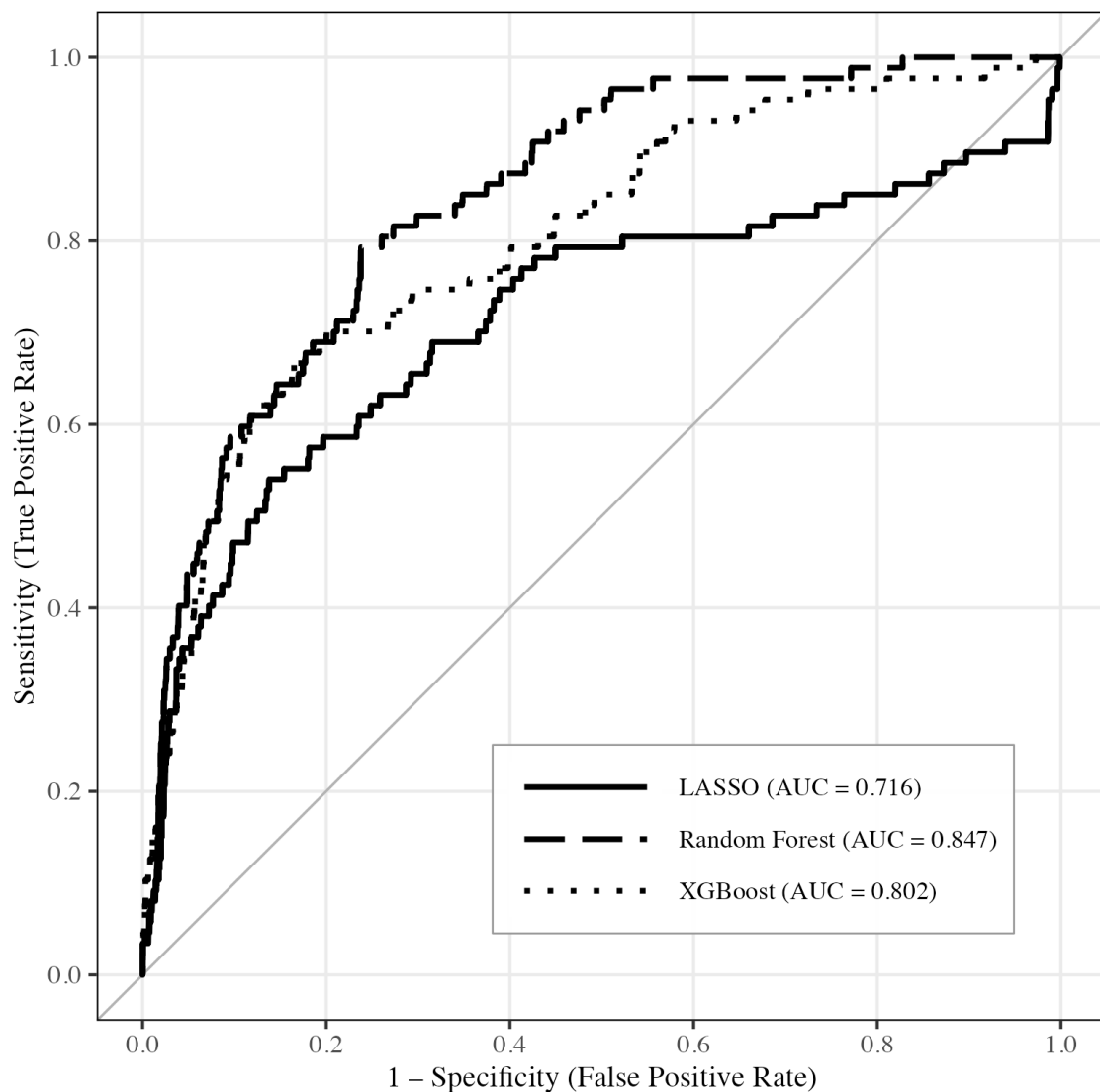
Model Comparison

Table 1 summarizes the performance of the three models on the held-out test set (2024Q1–2025Q3). To assess the statistical uncertainty beyond AUC, performance metrics were interpreted in light of the small number of positive cases ($n = 29$), which inherently increases variability in threshold-dependent measures such as precision and F1 score. This limitation underscores the importance of complementary metrics and temporal validation when evaluating rare-event prediction models. The random forest achieved the highest discrimination, with an AUC of 0.847 (95% CI: 0.809–0.885), followed by XGBoost at 0.802 (0.752–0.853) and LASSO at 0.716 (0.646–0.786). Pairwise DeLong tests confirmed that both ensemble models significantly outperformed LASSO (RF vs. LASSO: $p < 0.001$; XGBoost vs. LASSO: $p = 0.002$). The confidence intervals of RF and XGBoost partially overlapped (0.809–0.885 vs. 0.752–0.853). The ROC curves for all three models are displayed in Figure 1.

Table 1. Model performance comparison on temporal test set (2024–2025).

Model	AUC (95% CI)	Sens.	Spec.	Prec.	F1	Brier	N feat.
LASSO	0.716 (0.646–0.786)	0.540	0.862	0.078	0.137	0.020	11
Random Forest	0.847 (0.809–0.885)	0.793	0.762	0.067	0.124	0.020	77
XGBoost	0.802 (0.752–0.853)	0.621	0.881	0.102	0.174	0.026	77

Note: AUC = area under the receiver operating characteristic curve; Sens. = sensitivity; Spec. = specificity; Prec. = precision; N feat. = number of features. Best model in bold. Threshold selected by maximizing Youden’s J statistic. DeLong tests: LASSO vs. XGBoost, $p = 0.002$; LASSO vs. RF, $p < 0.001$.



Source: Elaborated by the authors based on ANS/DIOPS data (2018–2025).
Temporal validation: training $\leq 2023Q3$; test $\geq 2024Q1$ ($n = 4,107$).

Figure 1. ROC for the three predictive models on the temporal test set (2023–2025).

Source: Authors' elaboration.

Note: LASSO (dashed, AUC = 0.716), Random Forest (solid, AUC = 0.847), XGBoost (dotted, AUC = 0.802). Shaded areas = bootstrap 95% CI. DeLong: LASSO vs. RF, $p < 0.001$.

However, the generalization gap between training and test AUC revealed a starkly different picture. LASSO exhibited the smallest gap (0.014), indicating near-perfect generalization from training to test data (AUC = 0.730). The random forest showed a moderate gap of 0.119 (training AUC = 0.966), and XGBoost displayed the largest gap of 0.197 (training AUC = 0.999), suggesting substantial overfitting despite regularization. This large generalization gap indicates that XGBoost captured training-specific patterns that did not fully generalize to future observations. Such behavior is characteristic of high-capacity models applied to limited rare-event datasets and reinforces the importance of

evaluating both discrimination and generalization stability when selecting models for regulatory early warning systems. Among the secondary metrics, XGBoost achieved the highest F1 score (0.174) and precision (0.102). In contrast, the random forest achieved the highest sensitivity (0.793) and the lowest Brier score (0.020), indicating superior probabilistic calibration relative to the other models. This suggests that the random forest's predicted probabilities are better aligned with observed event frequencies, supporting its suitability for probabilistic risk estimation in regulatory applications.

To assess calibration beyond the Brier score, we computed the expected-to-observed (E:O) ratio, calibration intercept and slope from logistic recalibration, and the Hosmer-Lemeshow test. Random Forest showed the best calibration (Brier skill = +0.054; calibration slope = 0.935; E:O ratio = 0.73). LASSO showed acceptable calibration (Brier skill = +0.025; slope = 1.225; E:O = 0.75). XGBoost exhibited poor calibration, with a negative Brier skill score (−0.235) and an E:O ratio of 2.09, indicating systematic overestimation of risk. All three models showed significant Hosmer-Lemeshow tests ($p < 0.001$), expected given the extreme class imbalance and the sensitivity of the HL test to large samples (Figure 6).

The expanding-window cross-validation was used to assess the stability of LASSO over time. Across 16 temporal folds (test quarters from 2019Q3 through 2023Q3), the median AUC was 0.731 (IQR: 0.680–0.843; range: 0.548–0.944). Performance was notably lower in the initial COVID-19 quarter (2020Q1: AUC = 0.548) and peaked during the quarter with the highest concentration of distress events (2021Q4: AUC = 0.944). The number of features retained by the one-standard-error rule varied across folds (median: 10; range: 2–16), reflecting regularization's sensitivity to the evolving training composition. Complete fold-by-fold results are available in the code repository.

Table 2. Illustrative operator (ANS registry anonymized) retrospective backtest, selected quarters 2020–2025.

Quarter	P(LASSO)	P(XGB)	Pctile	N ops	Loss ratio	Comb. ratio	Op. margin
2020Q1	0.020	0.002	37.2	716	0.685	0.838	0.163
2020Q4	0.016	0.003	27.2	879	0.689	0.843	0.157
2021Q2	0.019	0.010	59.0	712	0.717	0.883	0.117
2021Q4	0.017	0.107	72.8	389	0.733	0.953	0.047
2022Q1	0.018	0.154	75.1	702	0.765	0.988	0.012

2022Q2	0.016	0.172	54.3	700	0.748	0.969	0.032
2023Q1	0.013	0.037	43.4	696	0.682	0.920	0.080
2023Q3	0.014	0.058	59.1	687	0.741	0.962	0.038
2024Q1	0.013	0.071	36.7	683	0.627	0.849	0.151
2024Q2	0.014	0.117	33.0	688	0.656	0.904	0.096
2025Q1	0.014	0.090	31.6	683	0.639	0.865	0.135
2025Q2	0.016	0.041	80.7	679	0.863	0.990	0.010
2025Q3	0.014	0.054	6.5	681	0.657	0.851	0.150

Note: P(LASSO) and P(XGB) = predicted distress probability. Pctile = percentile rank among all operators. Loss ratio, combined ratio, and operating margin are deaccumulated quarterly values. Bold = peak risk signal (2025Q2).

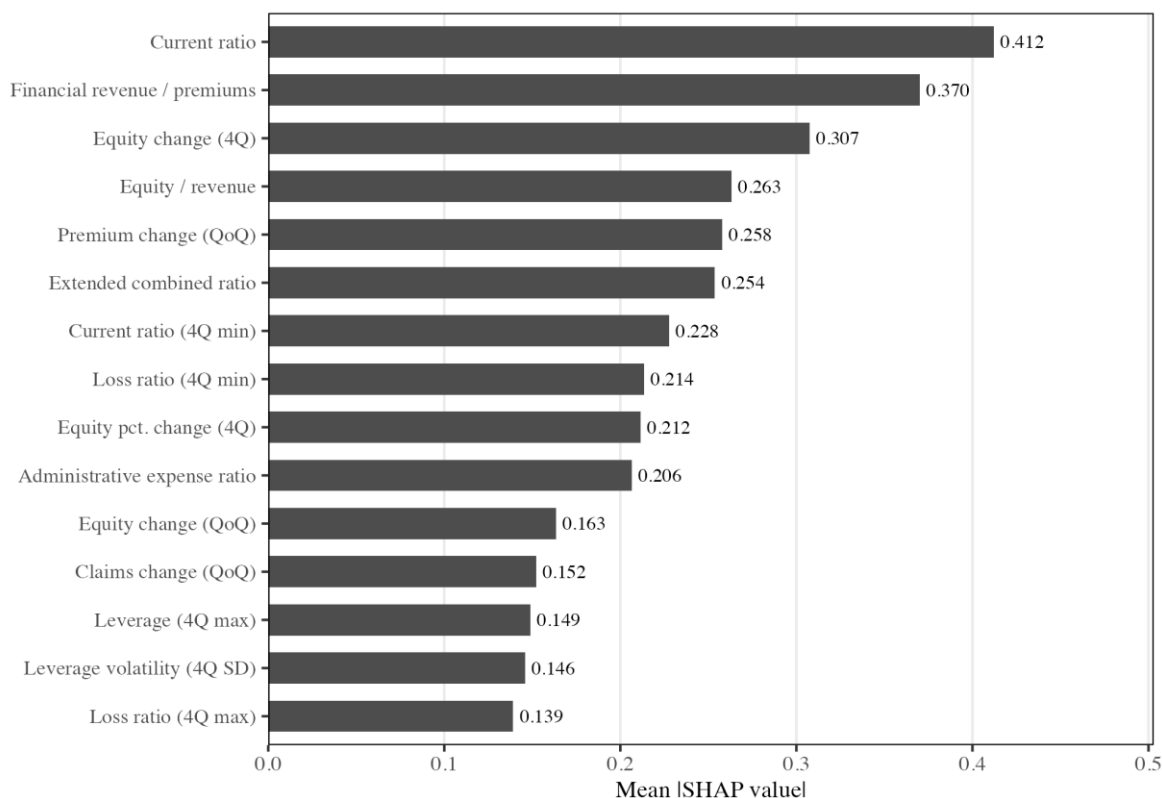
Feature Importance and Interpretability

Figure 2 presents the SHAP summary plot for the XGBoost model. The three features with the highest mean absolute SHAP values were: current ratio (liquidez corrente; 0.412), financial revenue-to-premiums ratio (receita financeira sobre contraprestações; 0.370), and four-quarter change in equity (variação do patrimônio líquido; 0.307). These were followed by the equity-to-revenue ratio (0.263) and quarter-over-quarter change in premiums (0.258). Notably, the top ten SHAP features collectively span liquidity (2 features), equity dynamics (3), profitability and cost structure (3), and revenue composition (2), indicating that the model draws on a broad set of financial dimensions rather than relying on a single indicator category.

The current ratio (SHAP rank 1, $|\text{SHAP}| = 0.412$) directly measures the operator's capacity to meet short-term obligations; operators with current ratio below 1.0 cannot cover monthly claims with current assets, signaling immediate liquidity risk. The financial revenue-to-premiums ratio (SHAP rank 2, $|\text{SHAP}| = 0.370$) proxies dependence on investment returns to compensate for negative operating margins, indicating a structurally unprofitable operation that becomes unsustainable when interest rates decline.

The four-quarter equity change (SHAP rank 3, $|\text{SHAP}| = 0.307$) captures accelerated equity deterioration—a measure of capital erosion velocity. The equity-to-revenue ratio (SHAP rank 4, $|\text{SHAP}| = 0.263$) reflects the relative capital buffer against adverse shocks. The extended combined ratio (LASSO rank 1, SHAP rank 6, $|\text{SHAP}| =$

0.254), the only consensus predictor, captures total costs relative to revenue; values above 1.0 indicate the operator spends more than it collects.



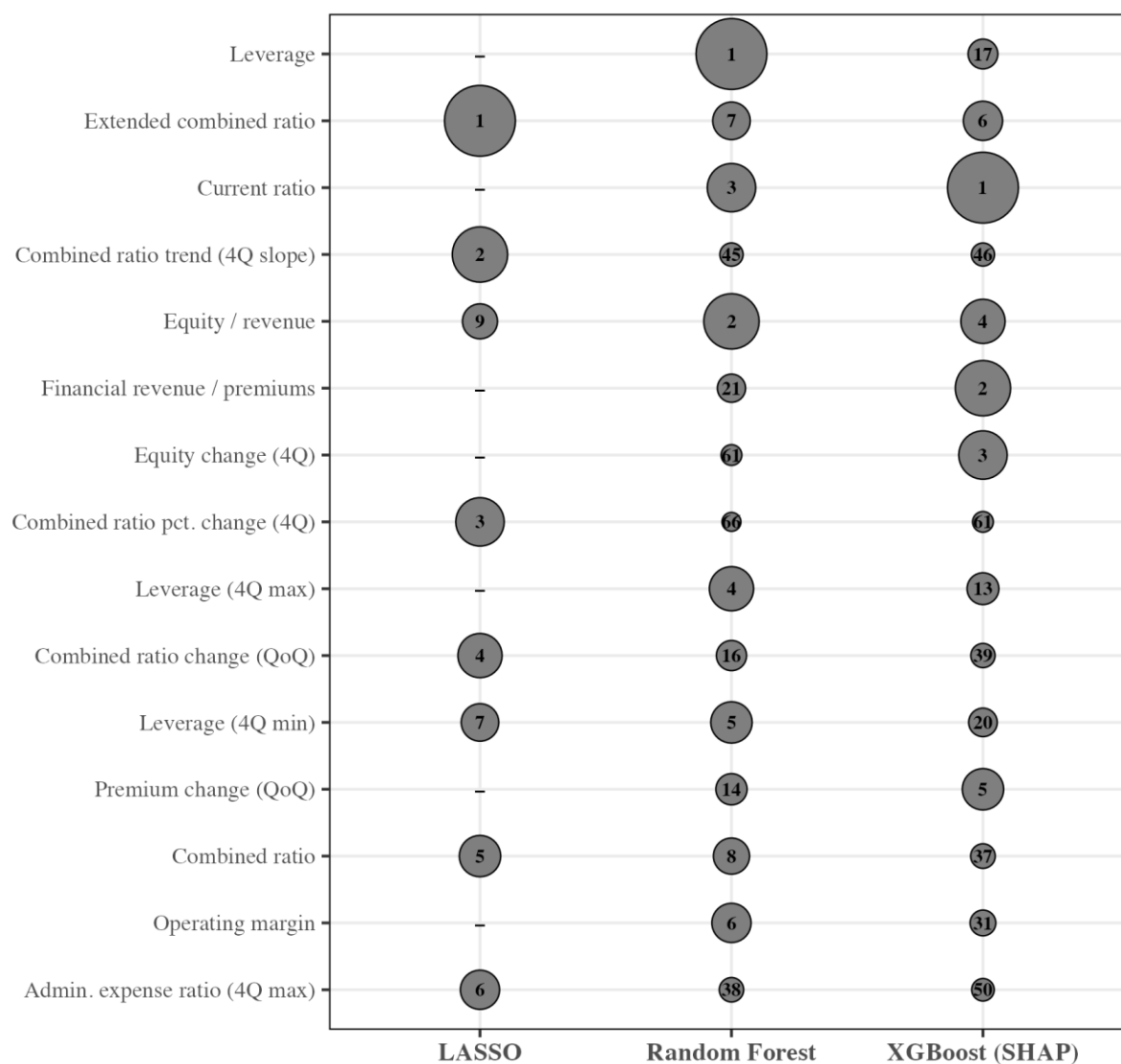
Source: Elaborated by the authors. SHAP values computed from the XGBoost model on the temporal test set ($n = 2,000$ observations).

Figure 2. SHAP summary plot for the XGBoost model (top 20 features)

Source: Authors' elaboration.

Note: Each point = one operator-quarter. Horizontal position = SHAP contribution to log-odds; color = feature value (red = high, blue = low).

Table 4 compares feature-importance rankings across the three models. The extended combined ratio (índice combinado ampliado) was the only feature to rank among the top seven across all three approaches (SHAP rank 6, RF rank 7, LASSO rank 1, with the highest coefficient of 0.289), making it the most robust consensus predictor of financial distress. Leverage (alavancagem) ranked first in the random forest but did not enter the LASSO model, a pattern that may reflect collinearity with other retained variables. Equity-to-revenue ratio (patrimônio líquido sobre receita) showed strong concordance across the ensemble models (SHAP rank 4, RF rank 2), though it carried less weight in the LASSO model (rank 9).



Source: Elaborated by the authors. Numbers indicate feature rank within each model. "-" = feature not selected (LASSO sparsity). Bubble size is inversely proportional to rank.

Figure 3. LASSO logistic regression coefficient path and selected features.

Source: Authors' elaboration.

Note: Panel A: regularization path; vertical dashed line = optimal λ (10-fold CV). Panel B: odds ratios with 95% CI for 11 retained features.

The LASSO model retained 11 of 77 candidate features (Figure 4, Table 3). Five of these were directly related to the combined ratio (level, slope, four-quarter percentage change, quarter-over-quarter change, and extended combined ratio), collectively accounting for the largest share of the model's predictive capacity. The two features with negative coefficients—minimum leverage over four quarters (OR = 0.972) and equity-to-revenue ratio (OR = 0.993)—are consistent with the expected direction: higher minimum leverage and stronger equity positions are protective against distress. Two features had coefficients very near zero (operating margin slope and quarter-over-quarter operating

margin change), indicating that the one-standard-error rule retained them at marginal regularization levels.

Table 3. Features selected by LASSO logistic regression (11 of 77 candidates).

Feature	Coefficient	Odds Ratio
Extended combined ratio	0.289	1.335
Combined ratio 4Q slope	0.161	1.175
Combined ratio 4Q % change	0.120	1.128
Combined ratio 1Q change	0.086	1.090
Combined ratio (level)	0.047	1.048
Administrative ratio 4Q max	0.036	1.037
Leverage 4Q range	0.023	1.024
Leverage 4Q min	-0.028	0.972
Equity-to-revenue ratio	-0.007	0.993
Operating margin 1Q change	< 0.001	1.000
Operating margin 4Q slope	< 0.001	1.000

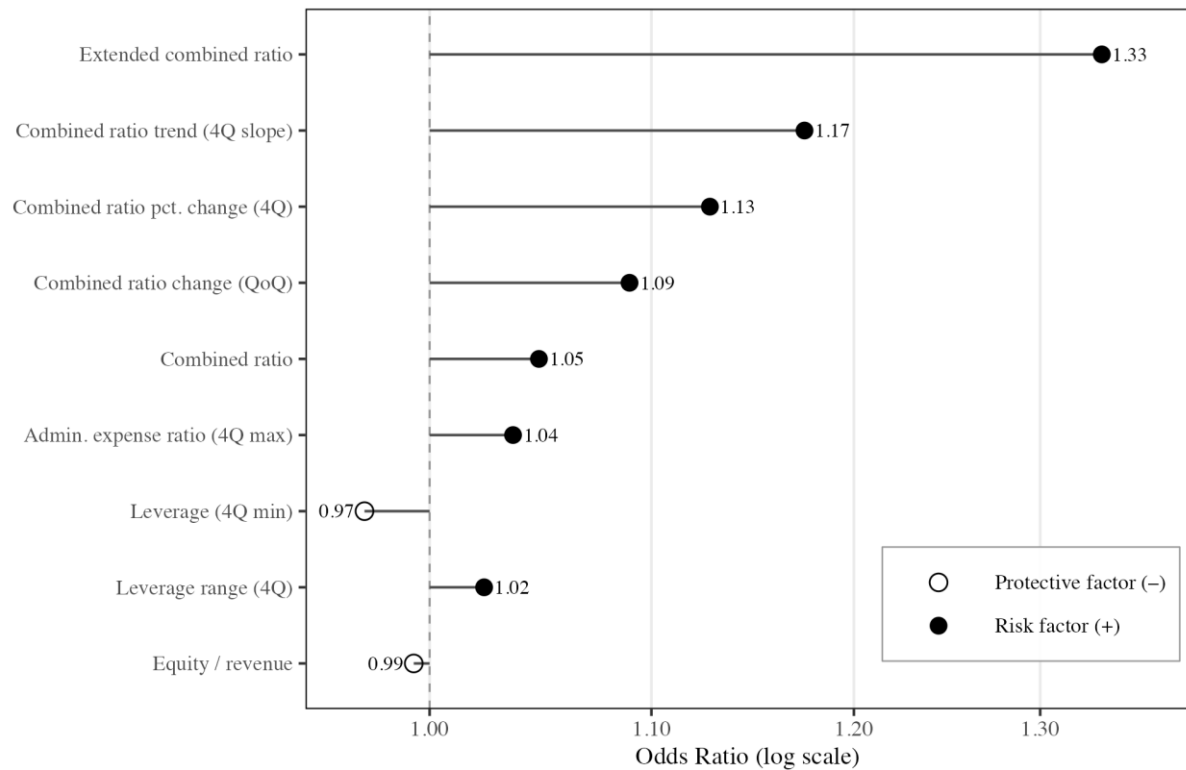
Note: Intercept = -4.568 (baseline odds = 0.010). OR = odds ratio. Features sorted by absolute coefficient magnitude.

Table 4. Top 15 features by SHAP importance (XGBoost) with cross-method comparison.

Feature	SHAP Rank	SHAP	RF Rank	LASSO Rank	Best Rank
Current liquidity	1	0.412	3	—	1
Financial income / premiums	2	0.370	21	—	2
Equity 4Q change	3	0.307	61	—	3
Equity-to-revenue ratio	4	0.263	2	9	2
Premiums 1Q change	5	0.258	14	—	5
Extended combined ratio	6	0.254	7	1	1
Current liquidity 4Q min	7	0.228	18	—	7
Loss ratio 4Q min	8	0.214	50	—	8
Equity 4Q % change	9	0.212	25	—	9
Administrative ratio	10	0.206	9	—	9
Equity 1Q change	11	0.163	47	—	11
Claims 1Q change	12	0.152	27	—	12
Leverage 4Q max	13	0.149	4	—	4
Leverage 4Q SD	14	0.146	13	—	13

Loss ratio 4Q max	15	0.139	49	—	15
-------------------	----	-------	----	---	----

Note: |SHAP| = mean absolute SHAP value. RF = Random Forest permutation importance rank. LASSO Rank among selected features (— = not selected). Best Rank = minimum across methods.



Source: Elaborated by the authors. LASSO logistic regression (λ at 1 SE rule).
 $OR > 1$ increases distress probability; $OR < 1$ is protective. 9 features selected from 77 candidates.

Figure 4. LASSO logistic regression coefficients for financial distress prediction (odds ratios).

Note: LASSO logistic regression (λ at 1 SE rule). $OR > 1$ increases distress probability; $OR < 1$ is protective. Eleven features retained from 77 candidates; nine with $|\text{coefficient}| > 0.005$ displayed. Two marginal features (operating margin 1Q change and 4Q slope) omitted.

Source: Authors' elaboration.

Among the temporal features, the four-quarter slope of the combined ratio (índice combinado slope4) was the second most important predictor in the LASSO model ($OR = 1.175$), indicating that trend-based features contributed predictive value beyond static ratio levels—a pattern consistent with Proposition P3, as discussed below.

Illustrative Case Study

The following analysis is presented as an illustrative case study rather than external validation. A single-operator backtest cannot substitute systematic prospective

evaluation; its purpose is to demonstrate the temporal behavior of model predictions under real-world conditions.

Figure 5 displays the expanding-window backtest for the selected operator from 2020Q1 through 2025Q3 (20 quarterly observations; Q4 values are unavailable due to the lag in DIOPS submission at the time of data extraction). The two models exhibited markedly different temporal profiles. LASSO produced highly stable predicted probabilities throughout the period, oscillating within a narrow band between 0.013 and 0.020—hovering around its classification threshold of 0.015 regardless of the operator's actual financial trajectory. This near-constant output meant that LASSO flagged the operator as distressed in roughly half of the quarters, without temporal correspondence to periods of genuine financial pressure. XGBoost, in contrast, showed pronounced sensitivity to the operator's financial trajectory. Predicted probabilities remained low through 2021Q1 ($p = 0.007$), then rose sharply to 0.107 in 2021Q4 as the operator's combined ratio exceeded 0.95. The XGBoost signal peaked at 0.171 in 2022Q2, during the period of most severe financial pressure—the operator's operating margin reached its nadir of 1.2% one quarter earlier (2022Q1). It is worth noting that this period coincided with broader structural pressures faced by many large operators, including accelerated medical cost inflation, rising health-related litigation, and, in the case of vertically integrated operators, the financial burden of rapid expansion through acquisitions of hospitals and clinics—factors that, while not directly incorporated as model features, contributed to the observed deterioration in financial ratios. A secondary elevation occurred in 2024Q2–Q3 ($p \approx 0.116$), coinciding with a combined ratio above 0.90 and an operating margin below 10%. By 2025Q2, despite a combined ratio of 0.990 and an operating margin of just 1.0%, the XGBoost signal had declined to 0.041—below the levels observed during the 2024Q2–Q3 elevation.

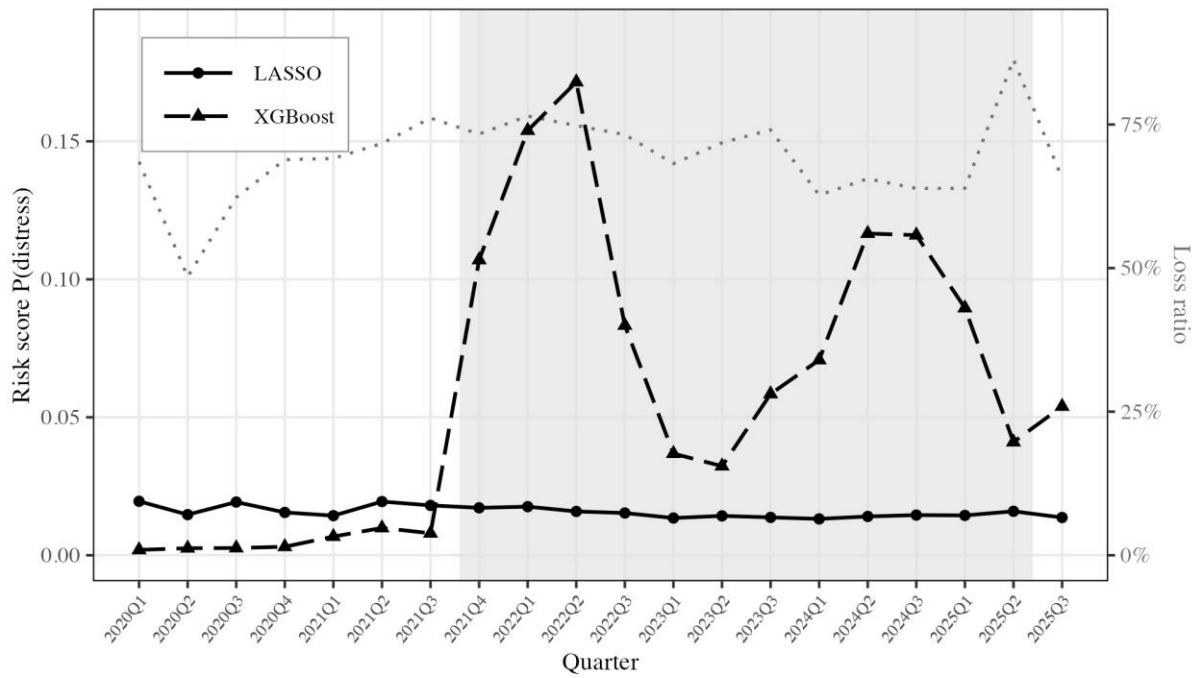


Figure 5. Illustrative operator retrospective backtest, 2020Q1–2025.

Source: Authors' elaboration.

Note: Panel A: XGBoost predicted probability (line) and percentile rank (bars); dashed line = 75th percentile. Panel B: deaccumulated quarterly loss ratio, combined ratio, and operating margin. Gray area = COVID-19 period.

The contrast between models is noteworthy. XGBoost first crossed its classification threshold in 2021Q4, one to two quarters before the period of lowest operating margin (2022Q1–Q2), and remained elevated through 2024Q3. LASSO's near-flat probability trajectory, by contrast, showed no temporal correspondence to the operator's financial deterioration. The XGBoost probability ranged from 0.002 to 0.171 over the backtest period, a degree of volatility consistent with the large generalization gap (0.197) observed in the overall evaluation.

To assess robustness to the missing value imputation strategy, we compared three approaches: zero imputation (original), complete exclusion of operators with fewer than eight quarters, and quarter-median imputation. The model ranking was preserved across all strategies (RF > XGBoost > LASSO). For Random Forest, AUC ranged from 0.814 to 0.839 ($\Delta_{\max} = 0.025$); for XGBoost, from 0.785 to 0.793 ($\Delta_{\max} = 0.008$); for LASSO, from 0.632 to 0.677 ($\Delta_{\max} = 0.045$).

DISCUSSION

The most consequential finding of this study is not which algorithm achieves the highest AUC. However, that predictive discrimination and generalization stability diverge sharply—and that this divergence carries direct implications for the institutional design of health insurance supervision. The test-set hierarchy—random forest (AUC = 0.847), XGBoost (0.802), LASSO (0.716)—favors ensemble methods, yet LASSO exhibited the smallest gap between training and test performance (0.014 vs. 0.119 for RF and 0.197 for XGBoost). For a regulatory agency that must balance the cost of false alarms against the systemic risk of missed signals, the choice between these models is not a statistical question. It is an institutional one. We organize the discussion around this central tension, examining how it intersects with the literature on financial distress prediction, feature dynamics, and the design of interpretable regulatory tools.

The low precision across all models (maximum 0.102 for XGBoost) warrants discussion of practical utility. The EWS is designed as a screening tool, not a standalone decision rule: from approximately 700 active operators, the model narrows the surveillance set to roughly 70 for in-depth inspection. The marginal cost of inspecting false positives is negligible compared to the systemic cost of missing an operator in deterioration, a cost asymmetry that justifies high-sensitivity models even at low precision. The classification threshold is adjustable to regulatory operational capacity.

The ensemble advantage reproduces a well-documented pattern. Barboza et al. (2017) reported that ensemble methods outperformed logistic regression by approximately 10 percentage points; Lessmann et al. (2015) confirmed this across 41 international datasets; and Dasilas and Rigani's (2024) review of 232 studies identified random forest and gradient boosting as consistently top-ranked. Our results support Proposition P1 and extend these findings to a domain with structurally distinct characteristics: the target is not terminal insolvency but heterogeneous regulatory sanctions; distress prevalence is extremely low; and reporting is limited to quarterly frequency. These features place health insurance closer to the tasks studied by Eling and Pankoke (2016) and Brockett et al. (1994)—who reported lower AUCs for insurer solvency—than to corporate bankruptcy benchmarks. The random forest AUC of 0.847 compares favorably with the 0.78–0.86 range in insurer-specific models. To our knowledge, this is the first application of ensemble ML to health insurance solvency prediction using a national regulatory database.

However, the AUC hierarchy tells only part of the story. The near-perfect training AUC of XGBoost (0.999) does not indicate superior learning—it indicates memorization.

Shumway (2001) argued that temporal out-of-sample validation is the only credible test for distress models. By this criterion, LASSO—despite ranking last in test-set AUC—emerges as the most trustworthy model: its 0.014 gap indicates that what it learns, it retains. The random forest occupies an intermediate position (gap = 0.119). This three-way pattern—high discrimination with high variance, moderate with moderate, low with low—complicates the common framing that presents ensemble methods as unambiguous improvements. In regulatory contexts with rare events, the “best” model depends on the decision environment, not merely the test-set metric.

The feature importance analysis supports Proposition P2 but also qualifies it. The extended combined ratio was the only consensus predictor ranked among the top seven features across all three methods (SHAP rank 6, RF rank 7, LASSO rank 1)—the most robust cross-model indicator of impending distress. This extends the finding of Cummins et al. (1999), who identified the combined ratio as a primary driver of U.S. insurer insolvency, and corroborates the Brazilian evidence of Bragança et al. (2019) and Reis et al. (2021). However, the SHAP analysis reveals that the ensemble does not collapse into a single-ratio proxy. The top ten XGBoost features span liquidity, equity dynamics, profitability, and revenue composition—a multidimensional risk profile that vindicates the theoretical argument, dating back to Altman (1968) and Ohlson (1980), that financial distress is irreducibly multicausal. The combined ratio is necessary but insufficient. The practical consequence is that composite surveillance models are likely to outperform the single-ratio threshold triggers currently embedded in the IDSS framework under Resolução Normativa nº 510 (ANS, 2022)—a design that our results suggest may be structurally unable to capture the full dimensionality of operator risk.

Temporal features support Proposition P3, strengthening the dynamic perspective Shumway (2001) advocated. Five of eleven LASSO-retained features were trend-based, with the four-quarter combined ratio slope as the second most important predictor (OR = 1.175)—extending to health insurance what Cole and Gunther (1998) demonstrated for bank failure. The illustrative case study made this concrete: XGBoost detected rising distress one to two quarters before operating margin reached its nadir, within the lead-time window reported by Sahajwala and Van den Bergh (2000) for banking EWS. However, by 2025Q2, despite a combined ratio of 0.990, the XGBoost probability had declined to 0.041. The model ceased to distinguish chronic stress from the operator’s new baseline—a pattern resembling concept drift (Gama et al., 2014). In an EWS context, this

represents a particularly insidious failure mode: the regulator is reassured precisely when vigilance matters most. This phenomenon warrants dedicated investigation.

Rudin (2019) argued that high-stakes decisions should rely on inherently interpretable models. Our results partially vindicate this: LASSO—transparent, 11 features, near-zero generalization gap—performs well enough as an automated screening tool. For narrowing surveillance from hundreds of operators to a manageable watchlist, parsimony is the performance. However, SHAP analysis reveals risk dimensions that LASSO cannot capture: nonlinear interactions among liquidity, equity, and revenue features. A regulator relying exclusively on LASSO would have a transparent but incomplete picture. We propose a two-tier architecture—LASSO screens and SHAP-interpreted ensemble investigations—paralleling the layered logic that Sahajwala and Van den Bergh (2000) described for banking supervision. This design treats interpretability as a continuum calibrated to decision stakes at each supervisory stage.

Taken together, these findings reframe the problem of health insurance supervision. The current IDSS framework relies on static ratio thresholds applied uniformly to all operators—an architecture that, as our results suggest, neither captures the multidimensionality of risk (P2) nor the temporal dynamics of deterioration (P3). A machine learning–based EWS built on publicly available DIOPS data offers a path toward dynamic, data-driven surveillance that any stakeholder can replicate and audit. In a sector where information asymmetry between regulators and operators has been identified as a structural vulnerability (Ocké-Reis, 2012; Trettel et al., 2018), open-data replicability is not merely a methodological convenience—it is a governance instrument. If the regulator's models can be independently verified, the asymmetry shrinks not through more data, but through more transparency.

Several limitations qualify these conclusions. First, precision was low across all models (maximum 0.102 for XGBoost), meaning roughly 9 out of 10 positive alerts would be false positives—a structural property of extreme class imbalance implying the EWS is suited as a screening tool, not a standalone decision rule. Second, quarterly DIOPS reporting constrains timeliness. Third, the single-country scope limits generalizability. Fourth, the exclusively financial feature set omits operational signals—such as beneficiary complaints, network adequacy, and claims denial rates—that may carry additional predictive power. Moreover, external factors such as investor expectations, management credibility, and strategic decisions (e.g., debt-financed vertical expansion) are not captured by the model's financial ratios, which limits its ability to fully

contextualize certain episodes of acute distress. Fifth, the illustrative backtest represents a single case rather than a systematic prospective evaluation; future validation should encompass multiple operators across different distress trajectories. Sixth, the XGBoost generalization gap (0.197) underscores the persistent risk of overfitting in rare-event settings. Future research should explore integrating textual data from regulatory proceedings, cost-sensitive recalibration, and prospective validation of the concept drift phenomenon.

FINAL CONSIDERATIONS

This study developed and validated the first machine-learning-based early-warning system for financial distress among Brazilian health plan operators. Its central message is that predicting distress and deploying prediction for regulation are fundamentally different problems. The models we tested differ not only in accuracy but also in the type of confidence they offer: ensemble methods provide more accurate estimates, while penalized regression generalizes more reliably. Choosing between them is an institutional decision, not a statistical one.

Three contributions follow from this finding. Theoretically, the study extends the EWS literature—rooted in banking and general insurance—to a domain characterized by extreme class imbalance, quarterly public data, and a regulatory target that precedes insolvency. Methodologically, the cross-model interpretability analysis demonstrates that financial distress in health operators is multidimensional and dynamic, requiring composite surveillance rather than single-ratio thresholds. In practice, the exclusive reliance on open regulatory data means the system can be independently replicated and audited, shifting surveillance from an opaque institutional process to a verifiable public function.

For the ANS and similar regulators, the implication is that static threshold frameworks can be complemented—not replaced—by data-driven screening tools that make risk dimensions and temporal trajectories visible. For operators, the results underscore that financial deterioration becomes legible to external models well before regulatory action materializes, creating both an opportunity and a responsibility to self-correct.

These conclusions are bounded by the low precision inherent to rare-event prediction, the single-country scope, and an exclusively financial feature set. Future research should address the model habituation phenomenon observed in the case study,

integrate operational and clinical variables, and conduct prospective validation as the regulatory data pipeline evolves. The tools exist. The open question is whether the institutions that govern health insurance are willing to let them reshape how supervision is practiced.

References

1. Agência Nacional de Saúde Suplementar. (2012). *Resolução Normativa nº 316, de 30 de novembro de 2012*. Diário Oficial da União. <https://www.gov.br/ans/pt-br/assuntos/legislacao/resolucoes-normativas>
2. Agência Nacional de Saúde Suplementar. (2019). *Resolução Normativa nº 443, de 25 de janeiro de 2019*. Diário Oficial da União. <https://www.gov.br/ans/pt-br/assuntos/legislacao/resolucoes-normativas>
3. Agência Nacional de Saúde Suplementar. (2022). *Resolução Normativa nº 510, de 30 de março de 2022*. Diário Oficial da União. <https://www.gov.br/ans/pt-br/assuntos/legislacao/resolucoes-normativas>
4. Agência Nacional de Saúde Suplementar. (2024). *Painel de indicadores econômico-financeiros da saúde suplementar*. ANS. <https://www.gov.br/ans/pt-br/aceso-a-informacao/perfil-do-setor/dados-e-indicadores-do-setor>
5. Agência Nacional de Saúde Suplementar. (2025). *Dados do setor de saúde suplementar*. ANS. <https://www.gov.br/ans/pt-br/aceso-a-informacao/perfil-do-setor/dados-gerais>
6. Altman, E. 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>
7. Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, 83, 405–417. <https://doi.org/10.1016/j.eswa.2017.04.006>
8. Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, 4, 71–111. <https://doi.org/10.2307/2490171>
9. Bragança, C. G., Pinheiro, L. E. T., Bressan, V. G. F., & Soares, L. A. C. F. (2019). Liquidação de operadoras de planos de assistência à saúde no Brasil. *Enfoque: Reflexão Contábil*, 38(2), 1–17. <https://doi.org/10.4025/enfoque.v38i2.43515>
10. Brasil. (1998). *Lei nº 9.656, de 3 de junho de 1998*. Dispõe sobre os planos e seguros privados de assistência à saúde. Diário Oficial da União. https://www.planalto.gov.br/ccivil_03/leis/19656.htm
11. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
12. Brockett, P. L., Cooper, W. W., Golden, L. L., & Pitaktong, U. (1994). A neural network method for obtaining an early warning of insurer insolvency. *The Journal of Risk and Insurance*, 61(3), 402–424. <https://doi.org/10.2307/253638>
13. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge*

- Discovery and Data Mining* (pp. 785–794). ACM. <https://doi.org/10.1145/2939672.2939785>
14. Cole, R. A., & Gunther, J. W. (1998). Predicting bank failures: A comparison of on- and off-site monitoring systems. *Journal of Financial Services Research*, 13(2), 103–117. <https://doi.org/10.1023/A:1007954718966>
 15. Conselho Nacional de Saúde. (2016). *Resolução nº 510, de 7 de abril de 2016*. Diário Oficial da União. <https://conselho.saude.gov.br/resolucoes/2016/Reso510.pdf>
 16. Cummins, J. D., Grace, M. F., & Phillips, R. D. (1999). Regulatory solvency prediction in property-liability insurance: Risk-based capital, audit ratios, and cash flow simulation. *The Journal of Risk and Insurance*, 66(3), 417–458. <https://doi.org/10.2307/253556>
 17. Dasilas, A., & Rigani, A. (2024). Machine learning techniques in bankruptcy prediction: A systematic literature review. *Expert Systems with Applications*, 255, 124761. <https://doi.org/10.1016/j.eswa.2024.124761>
 18. DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3), 837–845. <https://doi.org/10.2307/2531595>
 19. Eling, M., & Pankoke, D. (2016). Systemic risk in the insurance sector: A review and directions for future research. *Risk Management and Insurance Review*, 19(2), 249–284. <https://doi.org/10.1111/rmir.12062>
 20. Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4), Article 44, 1–37. <https://doi.org/10.1145/2523813>
 21. Instituto de Estudos de Saúde Suplementar. (2024). *Relatório IESS: Variação de custos médico-hospitalares 2024*. IESS. <https://www.iess.org.br/>
 22. Instituto de Estudos de Saúde Suplementar. (2025). *Judicialização na saúde suplementar: Desafios regulatórios e caminhos para a sustentabilidade do setor até 2035* (Série Caminhos da Saúde Suplementar, Ed. 10). IESS. <https://www.iess.org.br/>
 23. Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124–136. <https://doi.org/10.1016/j.ejor.2015.05.030>
 24. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* 30 (pp. 4765–4774). Curran Associates. <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>
 25. Ocké-Reis, C. O. (2012). *SUS: O desafio de ser único*. Editora Fiocruz.

26. Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18(1), 109–131. <https://doi.org/10.2307/2490395>
27. Reis, T. A., Macedo, M. A. S., & Marques, J. A. V. C. (2021). Desempenho econômico-financeiro e as decisões de instauração de regimes especiais no setor de saúde suplementar brasileiro. *Revista Contemporânea de Contabilidade*, 18(48), 156–174. <https://doi.org/10.5007/2175-8069.2021.e77327>
28. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
29. Sahajwala, R., & Van den Bergh, P. (2000). *Supervisory risk assessment and early warning systems* (Basel Committee on Banking Supervision Working Papers No. 4). Bank for International Settlements. https://www.bis.org/publ/bcbs_wp4.htm
30. Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The Journal of Business*, 74(1), 101–124. <https://doi.org/10.1086/209665>
31. Silva, V. V., & Loebel, E. (2016). Desempenho econômico-financeiro de operadoras de planos de saúde suplementar. *Revista de Administração Hospitalar e Inovação em Saúde*, 13(3), 1–17. <https://doi.org/10.21450/RAHIS.V13I3.3619>
32. Souza, A. A., & Silva, C. A. T. (2020). Desempenho econômico-financeiro e regulação no setor de saúde suplementar. *Revista de Administração Pública*, 54(4), 1–21. <https://doi.org/10.1590/0034-761220200121>
33. Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, 16(4), 437–450. [https://doi.org/10.1016/S0169-2070\(00\)00065-0](https://doi.org/10.1016/S0169-2070(00)00065-0)
34. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
35. Trettel, D. B., Kozan, J. F., & Scheffer, M. C. (2018). Judicialização em planos de saúde coletivos: Os efeitos da opção regulatória da Agência Nacional de Saúde Suplementar nos conflitos entre consumidores e operadoras. *Revista de Direito Sanitário*, 19(1), 166–187. <https://doi.org/10.11606/issn.2316-9044.v19i1p166-187>
36. Wang, D. W. L., Souza, F. M., Vasconcelos, N. P., Fajreldines, E., & Malik, A. M. (2024). O Judiciário aproxima o que a legislação separa: A judicialização da saúde no setor público e na saúde suplementar. *Revista de Direito Sanitário*, 24(1), e0011. <https://doi.org/10.11606/issn.2316-9044.rdisan.2024.213927>
37. Zhang, Z., Wu, C., Qu, S., & Chen, X. (2022). An explainable artificial intelligence approach for financial distress prediction. *Information Processing & Management*, 59(4), 102988. <https://doi.org/10.1016/j.ipm.2022.102988>