Machine learning early warning for financial distress in health plan operators

Guilherme Coelho, Clarimar José Coelho

https://doi.org/10.1590/SciELOPreprints.15181

Submitted on: 2026-02-22
Posted on: 2026-03-04 (version 1)
(YYYY-MM-DD)

MACHINE LEARNING EARLY WARNING FOR FINANCIAL DISTRESS IN HEALTH PLAN OPERATORS

Guilherme Coelho[1] · Clarimar José Coelho[2]

[1] Universidade Estadual de Campinas (UNICAMP), Faculdade de Ciências Médicas, Departamento de Tocoginecologia, Campinas, SP, Brazil

[2] Pontifícia Universidade Católica de Goiás (PUC Goiás), Escola de Ciências Exatas e da Computação, Goiânia, GO, Brazil

ORCID: G. Coelho — https://orcid.org/0000-0003-3045-4057 | C.J. Coelho — https://orcid.org/0000-0002-5163-2986

Corresponding author: Guilherme Coelho (guilcoel@unicamp.br)

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

Alerta temprana mediante aprendizaje automático para dificultades financieras en operadoras de planes de salud

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 on 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 dificuldades financeiras de operadoras de planos de saúde no Brasil. Com 24.440 observações operadora-trimestre de dados regulatórios públicos (2018–2025), validaram-se temporalmente 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), revelando tensão entre acurácia e confiabilidade regulatória. O índice combinado ampliado foi o único preditor consensual, e variáveis temporais agregaram valor preditivo além dos níveis estáticos. Estudo de caso na maior operadora do país demonstrou detecção precoce, mas também habituação do modelo sob estresse prolongado. O trabalho estende a teoria de sistemas de alerta do setor bancário à saúde suplementar e demonstra que dados públicos podem sustentar vigilância proativa e transparente onde a falência de operadoras compromete o acesso à saúde.

PALAVRAS-CHAVE Sistema de Alerta Antecipado; Dificuldade Financeira; Regulação de Planos de Saúde; Aprendizado de Máquina; Saúde Suplementar.

RESUMEN

Este estudio desarrolla el primer sistema de alerta temprana basado en aprendizaje automático para dificultades financieras de operadoras de planes de salud en Brasil. Con 24.440 observaciones operadora-trimestre de datos regulatorios públicos (2018–2025), se validaron temporalmente LASSO, random forest y XGBoost para predecir distress con dos a cuatro trimestres de anticipación. Random forest obtuvo mayor discriminación (AUC = 0,847), pero LASSO presentó menor brecha de generalización (0,014), revelando tensión entre precisión y confiabilidad regulatoria. El índice combinado ampliado fue el único predictor consensual, y variables temporales agregaron valor predictivo más allá de los niveles estáticos. Estudio de caso en la mayor operadora del país demostró detección temprana, pero también habituación del modelo bajo estrés prolongado. El trabajo extiende la teoría de alerta temprana del sector bancario al seguro de salud y demuestra que datos públicos pueden sustentar vigilancia proactiva donde la quiebra de operadoras compromete el acceso a la salud.

PALABRAS CLAVE Sistema de Alerta Temprana; Dificultad Financiera; Regulación de Seguros de Salud; Aprendizaje Automático; Salud Suplementaria.

INTRODUCTION

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). The dramatic case of Hapvida, Brazil's largest health insurer by enrollment, crystallized these vulnerabilities: in November 2025, the company lost R$ 7 billion in market capitalization in a single trading session after reporting third-quarter results significantly below expectations, including a cash-basis loss ratio of 75.2% and earnings before interest, taxes, depreciation and amortization (EBITDA) 20% below consensus estimates (Hapvida NotreDame Intermédica, 2025).

The ANS maintains a comprehensive prudential monitoring framework, including quarterly financial reporting through the disclosure of economic and financial information system (DIOPS), 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 Hapvida's deterioration before the November 2025 collapse? 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 been a central concern in accounting and finance since Beaver (1966) demonstrated that individual financial ratios—particularly cash flow to total debt—could discriminate between failing and non-failing firms up to five years before bankruptcy. Altman (1968) extended this univariate approach by combining five ratios into a single discriminant function, the Z-score, which became the foundational model for predicting corporate bankruptcy. Ohlson (1980) subsequently introduced logistic regression to the field, relaxing the restrictive distributional assumptions of discriminant analysis and producing probability estimates of failure rather than binary classifications. Together, these contributions established a statistical paradigm that would shape financial distress research for decades. A critical limitation of these classical models, however, is their reliance on cross-sectional snapshots of financial position: they classify firms at a single point in time rather than modeling deterioration trajectories. Shumway (2001) demonstrated that discrete-time hazard models incorporating time-varying covariates produced more accurate and consistent bankruptcy forecasts than static approaches, showing further that market-based variables complemented accounting ratios. This insight—that the temporal dynamics of financial indicators may be as informative as their levels—has important implications for EWS design in any regulated sector.

In banking regulation, EWS emerged as supervisory tools designed to identify vulnerable institutions before crises materialize. The CAMELS framework—evaluating Capital adequacy, Asset quality, Management competence, Earnings, Liquidity, and Sensitivity to market risk—became the standard off-site monitoring approach adopted by regulators worldwide (Sahajwala & Van den Bergh, 2000). Cole and Gunther (1998) showed that statistical models based on accounting ratios could match or outperform expert-based CAMELS ratings in predicting bank failures, establishing the empirical case for quantitative EWS. In the insurance sector, Cummins et al. (1999) developed solvency prediction models using risk-based capital ratios, audit indicators, and cash flow simulations, while Eling and Pankoke (2016) reviewed the growing body of evidence on systemic risk in insurance, noting that traditional actuarial indicators often fail to capture interconnected vulnerabilities. Despite this rich tradition in banking and insurance, no comparable EWS framework has been proposed for health insurance operators—a gap that is particularly consequential in regulated markets where operator failure directly affects access to healthcare services.

Machine Learning in Financial Prediction

The application of machine learning (ML) to financial prediction has challenged the dominance of traditional econometric approaches. Barboza et al. (2017) compared classical statistical models with ML algorithms across a large sample of U.S. firms and found that ensemble methods—particularly random forests and boosting—achieved AUC values above 0.87, substantially outperforming logistic regression and discriminant analysis. Similarly, Lessmann et al. (2015) benchmarked 41 classifiers across eight credit-scoring datasets and found that ensemble methods consistently ranked among the top performers, although the magnitude of improvement over well-tuned logistic regression varied by context. This performance gap is attributed to the ability of tree-based methods to capture nonlinear relationships and high-order interactions among financial ratios without requiring a priori specification of functional forms. For regulatory applications, however, predictive accuracy alone is insufficient: supervisors must justify intervention decisions with transparent evidence, creating a tension between model performance and interpretability that any operationally viable EWS must resolve.

The present study employs three algorithms that span the interpretability–performance spectrum. LASSO, logistic regression (Tibshirani, 1996) performs simultaneous estimation and variable selection by imposing an L1 penalty on coefficient magnitudes, shrinking irrelevant predictors to exactly zero. This property is particularly valuable in financial distress applications, where dozens of candidate ratios exhibit multicollinearity: LASSO identifies the most parsimonious subset of diagnostic indicators while producing interpretable coefficients with direct economic meaning. Random forests (Breiman, 2001) aggregate predictions from hundreds of decorrelated decision trees, each trained on bootstrap samples with random subsets of features, achieving low variance through ensemble averaging. Extreme gradient boosting (XGBoost) (Chen & Guestrin, 2016) implements gradient-boosted trees with regularization, sequentially fitting weak learners to the residuals of prior models. Both ensemble methods sacrifice direct coefficient interpretation for predictive gains, but this trade-off can be mitigated by post hoc interpretability tools.

Specifically, Lundberg and Lee (2017) proposed SHAP, grounded in cooperative game theory, which assigns each feature a consistent marginal contribution to individual predictions. SHAP values allow practitioners to decompose model outputs into feature-level effects, addressing the "black-box" criticism that has limited regulatory adoption of ML-based monitoring systems. The combination of a transparent linear model (LASSO) with opaque but

powerful ensembles (random forest, XGBoost), interpreted through SHAP, enables a rigorous comparison of the trade-off between parsimony and accuracy in the specific context of health insurer supervision.

The Brazilian Supplementary Health Sector

Brazil's supplementary health system is a hybrid arrangement in which private operators provide coverage to approximately 51 million beneficiaries under the regulatory oversight of the ANS (ANS, 2025). Operators submit quarterly financial statements through the documento de informações periódicas (DIOPS), which includes balance sheets, income statements, and operational data at the operator level—constituting a rich but underexploited panel dataset. The ANS monitors operators via composite indicators such as the IDSS, which aggregates financial, operational, healthcare quality, and beneficiary satisfaction dimensions into a single score. While the IDSS serves as a benchmarking tool, it was not designed as a predictive instrument: it assigns equal or arbitrary weights to its dimensions, does not model temporal dynamics, and produces an overall assessment rather than a probability of future distress. When financial deterioration reaches critical thresholds, the ANS may impose special regulatory regimes—fiscal direction, technical direction, or liquidation—under the provisions of regulatory resolution (Resolução Normativa) nº 510 (ANS, 2022). This regulatory architecture, however, operates primarily in a reactive mode: interventions typically occur after distress is already manifest (Ocké-Reis, 2012).

A small but growing literature has examined the financial performance of Brazilian health operators using accounting data. Silva and Loebel (2016) analyzed profitability and liquidity ratios across operator modalities from 2001 to 2012 using descriptive statistics and mean-comparison tests, identifying persistent differences among medical cooperatives, group medicine, and self-managed plans, but did not develop predictive models. Bragança et al. (2019) employed logistic regression with panel data and 26 financial variables to investigate the determinants of operator liquidation, concluding that leverage and provisioning ratios were the strongest predictors of failure; however, their model was estimated on a single cross-section of liquidated versus active firms and was not validated out of sample. Reis et al. (2021) extended this line of inquiry using factor analysis followed by logistic regression with resampling to predict ANS decisions to establish special regimes, finding that liquidity, leverage, and financial slack were significant determinants up to two years before intervention—the closest existing approximation to a predictive EWS, though still limited by

its reliance on a single linear method without temporal feature engineering. These studies demonstrate the relevance of financial ratios in the Brazilian context but share a common methodological limitation: they rely exclusively on retrospective descriptive analyses or traditional econometric methods.

The absence of forward-looking predictive models is consequential for two reasons. First, the regulatory tools currently available to the ANS are fundamentally backward-looking—composite indicators summarize past performance rather than estimating future risk trajectories. By the time an operator enters a special regime, financial deterioration has typically progressed to a point where recovery is difficult, and beneficiary harm has already occurred. Second, the sector's rapid consolidation has increased systemic interdependence: the failure of a large operator can cascade through provider networks, disrupt regional access to healthcare, and impose unfunded liabilities on the public system (ANS, 2024). A machine learning–based EWS, trained on the same publicly available DIOPS data that regulators already collect, could provide the anticipatory capacity that existing monitoring tools lack. By combining a transparent linear model with powerful ensemble methods—interpreted through SHAP—such a system would enable regulators to shift from reactive intervention to proactive risk management while maintaining the accountability required by the regulatory mandate. Drawing on the foregoing review, we derive three testable propositions that guide our empirical analysis:

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 on a quarterly basis. Our sample covers 30 quarters from the first quarter of 2018 (2018Q1) through the third quarter of 2025 (2025Q3), encompassing between 679 and 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-hospitalares), 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 but 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) designed to capture the range of regulatory outcomes that signal severe operational deterioration. Specifically, Y = 1 if any of the following events occurs within a forward-looking window of K = 4 quarters (one year): (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 regulatory resolution (Resolução Normativa) nº 316 (ANS, 2012), or (c) classification in prudential categories S3 or S4 (indicating heightened and critical risk, respectively) under the risk-based supervision framework established by Resolução Normativa nº 443 (ANS, 2019). This composite definition is preferable to single-event targets because operators under severe stress may experience any combination of these outcomes, and restricting the target to cancellation alone would miss cases where regulatory intervention prevents formal closure. The four-quarter horizon balances the utility of early detection—providing regulators and investors with sufficient lead time to act—against signal strength, as shorter horizons reduce the number of

observable events and longer horizons dilute the signal with noise. Operator-quarter observations occurring after the event are excluded from the analysis, as post-event financial data do not represent the predictive setting. In the training set, the prevalence of Y = 1 is approximately 1.7%, reflecting the rarity of severe distress events—a characteristic class imbalance that informs our modeling choices.

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 capturing deterioration dynamics: quarter-over-quarter changes (d_*_1q), 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 provided in Table 1). 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 sensitivity analyses excluding operators with fewer than eight quarters of data yielded substantively identical 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 wpos =

N0/N1 ≈ 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 = √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 (sensitivity + 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 (AUC_train–AUC_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 2 and visualized through ROC curves in Figure 1.

Interpretability and Case Study

To assess the practical value of the EWS, we conduct a retrospective case study on Hapvida Assistência Médica (REG_ANS 368253), Brazil's largest health plan operator by number of beneficiaries. Using an expanding-window backtest from 2020Q1 through 2025Q3, we retrain each model at every quarter and score only Hapvida, producing a time series of predicted distress probabilities. This design allows us to evaluate whether the models could have provided early warning signals of Hapvida's financial deterioration—a case of significant public interest given the operator's market share and the regulatory scrutiny it has attracted (Figure 4). Feature importance is compared across all three models (Table 3) 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 5.

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 https://github.com/guilcoel/ews-health-operators, 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 is inherent to the rarity of severe regulatory interventions and underscores the importance of confidence intervals and complementary validation approaches when interpreting model performance. Table 1 presents the distributions of the 77 candidate features and compares the training and test sets with respect to key financial indicators, including loss ratio, extended combined ratio, and current ratio.

[Insert Table 1 about here]

Model Comparison

Table 2 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.

[Insert Table 2 about here]

[Insert Figure 1 about here]

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 (AUC = 0.730) to test data. 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), whereas 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 predicted probabilities from the random forest are better aligned with observed event frequencies, supporting its suitability for probabilistic risk estimation in regulatory applications.

The expanding-window cross-validation was used to assess LASSO stability 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 the sensitivity of regularization to the evolving training composition. Table 4 presents the complete fold-by-fold results.

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.

[Insert Figure 2 about here]

Figure 3 compares feature-importance rankings across the three models. The extended combined ratio (índice combinado ampliado) was the only feature that ranked among the top seven in 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).

[Insert Figure 3 about here]

The LASSO model retained 11 of 77 candidate features (Figure 5, 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.

[Insert Figure 5 about here]

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.

Hapvida Case Study

Figure 4 displays the expanding-window backtest for Hapvida 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 Hapvida as distressed in roughly half of the quarters, without temporal correspondence to periods of genuine financial pressure. XGBoost, in contrast, showed pronounced sensitivity to Hapvida'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—Hapvida's operating margin reached its nadir of 1.2% one quarter earlier (2022Q1). 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.

[Insert Figure 4 about here]

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 Hapvida'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.

DISCUSSION

The most consequential finding of this study is not which algorithm achieves the highest AUC, but 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, the role of feature dynamics, and the design of interpretable regulatory tools.

The ensemble advantage observed here reproduces a pattern well documented in corporate bankruptcy prediction. Barboza et al. (2017) reported that ensemble methods outperformed logistic regression by approximately ten percentage points in accuracy; Lessmann et al. (2015) confirmed this margin across 41 international datasets; and the systematic review by Dasilas and Rigani (2024), covering 232 studies, identified random forest and gradient boosting as the most consistently top-ranked algorithms. Our results

provide direct support for Proposition P1 and extend these findings to a domain with structurally distinct characteristics: the prediction target is not terminal insolvency but a heterogeneous set of regulatory sanctions; distress prevalence is extremely low (1.75% in training, 0.44% in testing); and reporting is limited to quarterly frequency. These features place health insurance closer to the insurance-specific prediction tasks studied by Eling and Pankoke (2016) and Brockett et al. (1994)—who reported lower AUCs for insurer solvency than for corporate bankruptcy, owing to the opacity of insurance liabilities and the stochastic nature of claim reserves—than to the corporate bankruptcy benchmarks where AUCs above 0.90 are common. The random forest AUC of 0.847 compares favorably with the 0.78–0.86 range reported in insurer-specific models, suggesting that the ensemble approach remains effective even under these more demanding conditions. To our knowledge, this constitutes the first application of ensemble machine learning to health insurance solvency prediction using a national regulatory database.

Yet 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, precisely because future economic conditions may deviate from those in the training window. By this criterion, LASSO—despite ranking last in test-set AUC—emerges as the most trustworthy model: its gap of 0.014 means that what it learns, it retains. The random forest occupies an intermediate position (gap = 0.119), offering substantially better discrimination than LASSO while maintaining moderate stability. This three-way pattern—high discrimination with high variance, moderate discrimination with moderate variance, low discrimination with low variance—complicates the common framing in the machine learning literature, where ensemble methods are presented as unambiguous improvements over linear baselines (Dasilas & Rigani, 2024). In regulatory contexts with rare events and shifting economic conditions, the "best" model is a function of the decision environment, not merely of 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, present since 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.

The contribution of temporal features supports Proposition P3 and strengthens the dynamic perspective on financial distress that Shumway (2001) advocated for corporate settings. Five of the eleven features retained by LASSO were trend-based (slopes, ranges, quarter-over-quarter changes), and the four-quarter slope of the combined ratio was the second most important predictor (OR = 1.175)—extending to health insurance what Cole and Gunther (1998) demonstrated for bank failure: that trajectories predict what levels alone cannot. The Hapvida case study made this point concrete. XGBoost detected rising distress one to two quarters before the operator's operating margin reached its nadir, within the one-to-four-quarter lead-time window reported by Sahajwala and Van den Bergh (2000) for banking EWS. But the case also exposed a more subtle problem. By 2025Q2, despite a combined ratio of 0.990 and an operating margin of just 1.0%, the XGBoost probability had declined to 0.041. The model did not fail to detect stress—it ceased to distinguish chronic stress from the operator's new baseline. This pattern resembles what the machine learning literature terms concept drift: a shift in the data-generating process that renders learned decision boundaries obsolete (Gama et al., 2014). In an EWS context, it represents a particularly insidious failure mode—the regulator is reassured precisely when vigilance matters most. Whether this reflects a limitation of the four-quarter feature window, a property of gradient boosting under protracted stress, or a more general boundary condition of financial EWS remains an open empirical question. To our knowledge, it has not been examined in the distress prediction literature, and we suggest it warrants dedicated investigation.

Rudin (2019) argued that high-stakes decisions should rely on inherently interpretable models, not on post-hoc explanations of opaque ones. Our results partially vindicate and partially challenge this position. They vindicate it insofar as LASSO—fully transparent, 11 features with known coefficients, near-zero generalization gap—performs well enough to

serve as an automated screening tool. For the task of narrowing the surveillance perimeter from hundreds of operators to a manageable watchlist, interpretability is not sacrificed for performance; LASSO's parsimony is the performance. But they challenge Rudin's position insofar as the SHAP analysis (Lundberg & Lee, 2017) reveals risk dimensions that LASSO cannot capture: nonlinear interactions among liquidity, equity, and revenue features that only the ensemble detects. A regulator who relied exclusively on LASSO would have a transparent but incomplete picture. The resolution we propose—a two-tier architecture where LASSO screens and the SHAP-interpreted ensemble investigates—parallels the layered logic that Sahajwala and Van den Bergh (2000) described for banking supervision, where statistical screens trigger initial scrutiny and expert judgment provides deeper evaluation. This design does not treat interpretability as a binary choice but as a continuum calibrated to the decision stakes at each supervisory stage. To our knowledge, it represents the first application of SHAP-based interpretability to health insurance EWS, contributing to evidence that post-hoc explainability can reconcile accuracy and transparency in financial regulation (Zhang et al., 2022).

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, and most consequentially, precision was low across all models (maximum 0.102 for XGBoost), meaning roughly nine out of ten positive alerts would be false positives. This is not a modeling failure but a structural property of extreme class imbalance (1.75% positive cases); it implies that the EWS is suited as a screening tool to narrow the surveillance set, not as a standalone decision rule. Second, DIOPS quarterly reporting with submission lag constrains timeliness. Third, the single-country scope limits external generalizability. Fourth, the exclusively financial feature

set omits operational and clinical signals—beneficiary complaints, network adequacy, claims denial rates—that may carry additional predictive power (Silva & Loebel, 2016). Fifth, the Hapvida backtest, while analytically informative, represents a single case rather than a systematic prospective evaluation. Sixth, the XGBoost generalization gap (0.197) and the low prevalence of distress events (312 of 24,440 observations) underscore persistent overfitting risk in rare-event settings (Barboza et al., 2017). Future research should explore the integration of textual data from ANS regulatory proceedings, extension to operational risk dimensions, cost-sensitive recalibration that penalizes missed distress more heavily than false alarms, and—perhaps most urgently—prospective validation of the concept drift phenomenon observed in the Hapvida case.

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 in the type of confidence they offer: ensemble methods detect more, 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. Practically, the exclusive reliance on open regulatory data means the system can be independently replicated and audited, shifting surveillance from an opaque institutional process toward 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.

DATA AVAILABILITY STATEMENT

The analytical code is publicly available at https://github.com/guilcoel/ews-health-operators. The raw data used in this study are publicly available from the Brazilian National Supplementary Health Agency (ANS) through the DIOPS database (https://dados.gov.br/dados/conjuntos-dados/diops). Processed datasets in Parquet format are included in the repository.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS (CRediT)

Guilherme Coelho: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization, Project Administration.

Clarimar José Coelho: Methodology, Validation, Writing – Review & Editing, Supervision.

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

AI TOOLS DECLARATION

AI-assisted tools (Claude, Anthropic) were used for code review, manuscript formatting, and language editing. All analytical decisions, model specifications, interpretation of results, and scientific content were made entirely by the authors. The authors take full responsibility for the accuracy and integrity of the work.

REFERENCES

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

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

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

Agência Nacional de Saúde Suplementar. (2024). *Painel de indicadores econômico-financeiros da saúde suplementar*. ANS. https://www.gov.br/ans/pt-br/acesso-a-informacao/perfil-do-setor/dados-e-indicadores-do-setor

Agência Nacional de Saúde Suplementar. (2025). *Dados do setor de saúde suplementar*. ANS. https://www.gov.br/ans/pt-br/acesso-a-informacao/perfil-do-setor/dados-gerais

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

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

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, *4*, 71–111. https://doi.org/10.2307/2490171

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

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/l9656.htm

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

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

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

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

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

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

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

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

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

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

Hapvida NotreDame Intermédica. (2025). *Earnings release: Third quarter 2025 results*. Hapvida RI. https://ri.hapvida.com.br/

Instituto de Estudos de Saúde Suplementar. (2024). *Relatório IESS: Variação de custos médico-hospitalares 2024*. IESS. https://www.iess.org.br/

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.

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

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.

Ocké-Reis, C. O. (2012). *SUS: O desafio de ser único*. Editora Fiocruz.

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

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

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

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.

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

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

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

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

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

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

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

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

Table 1. 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 2. 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 optimized for F1. DeLong tests: LASSO vs. XGBoost, p = 0.002; LASSO vs. RF, p < 0.001.

Table 3. 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.

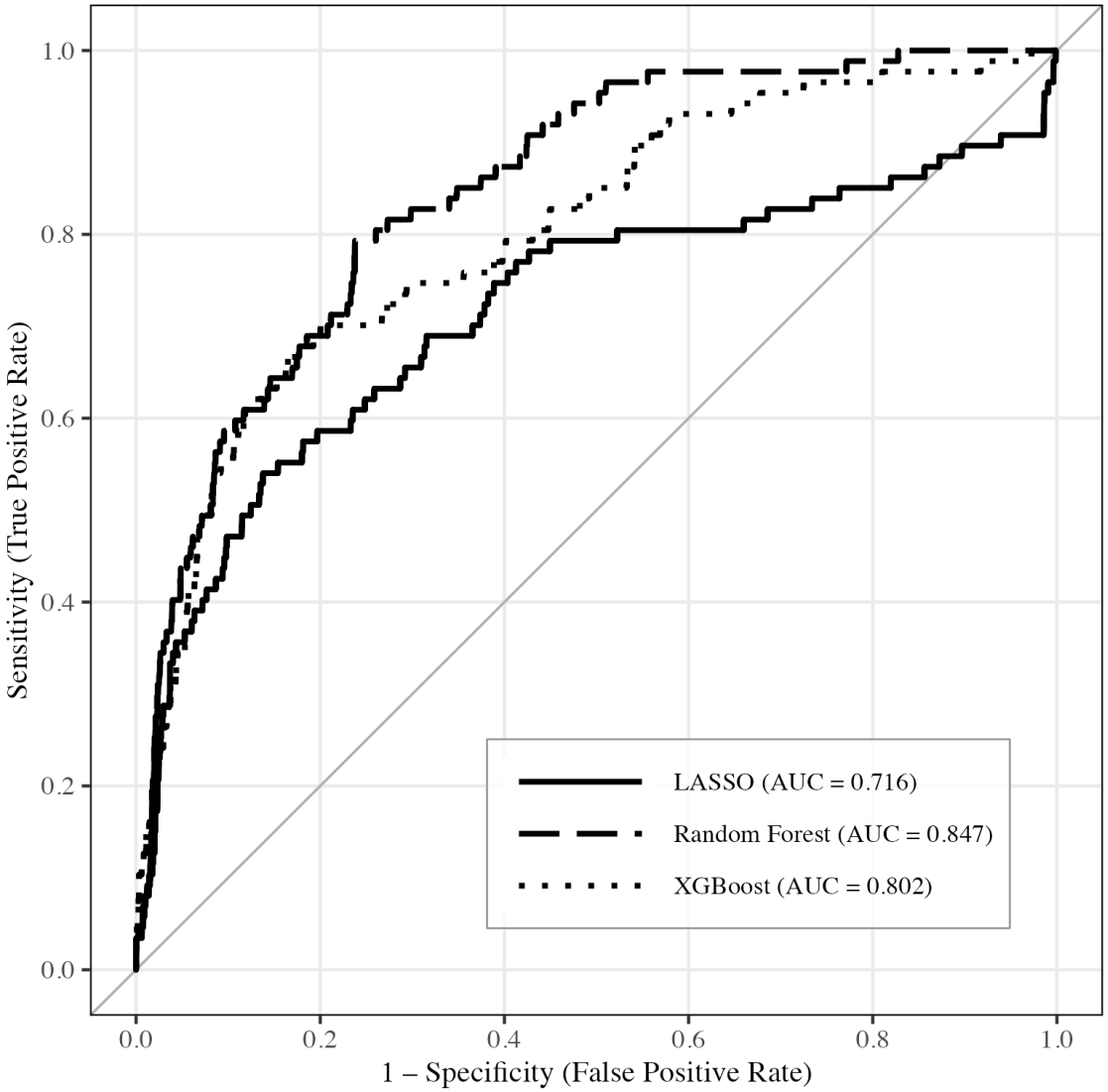
Table 4. Hapvida (ANS 368253) retrospective backtest, selected quarters 2020–2025.

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).

FIGURES

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

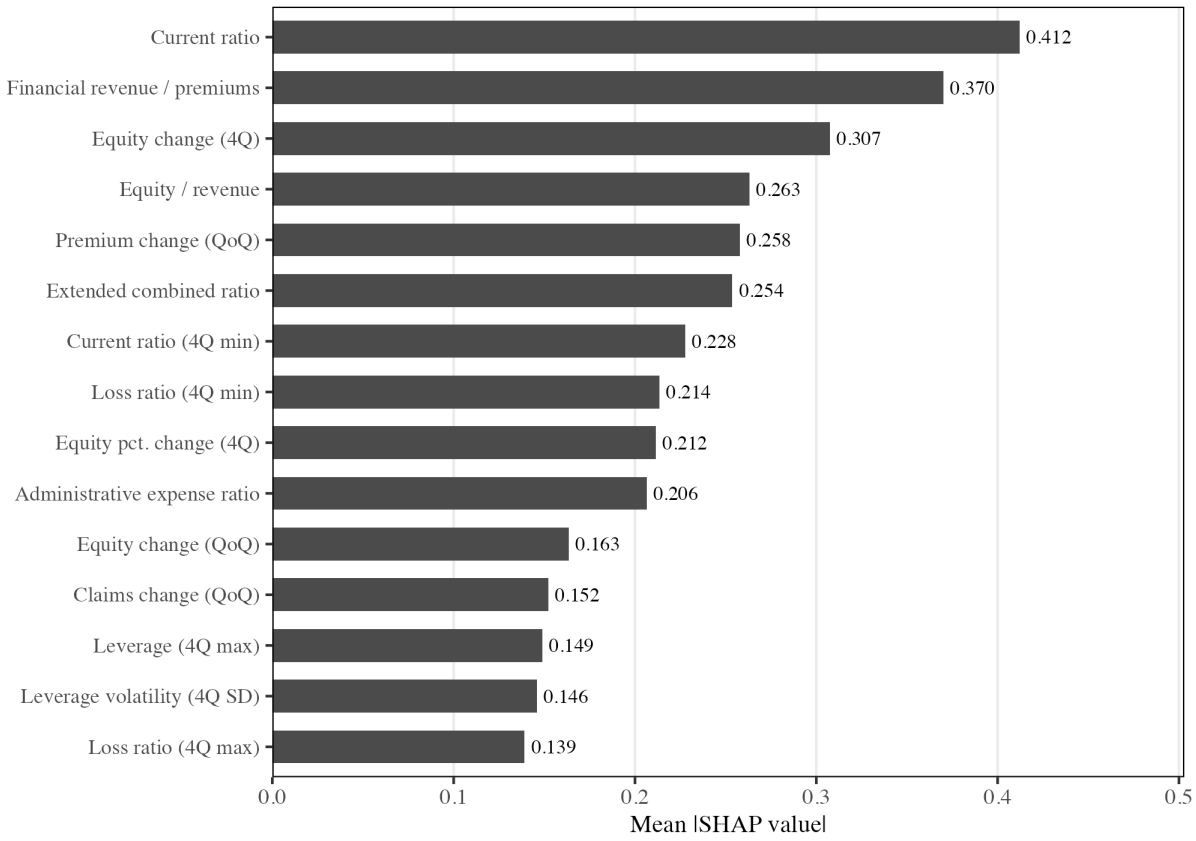


Source: Elaborated by the authors based on ANS/DIOPS data (2018–2025).
Temporal validation: training ≤ 2023Q3; test ≥ 2024Q1 (n = 4,107).

Source: Elaborated by the authors.

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, $p < 0.001$.

Figure 2. LASSO logistic regression coefficient path and selected features

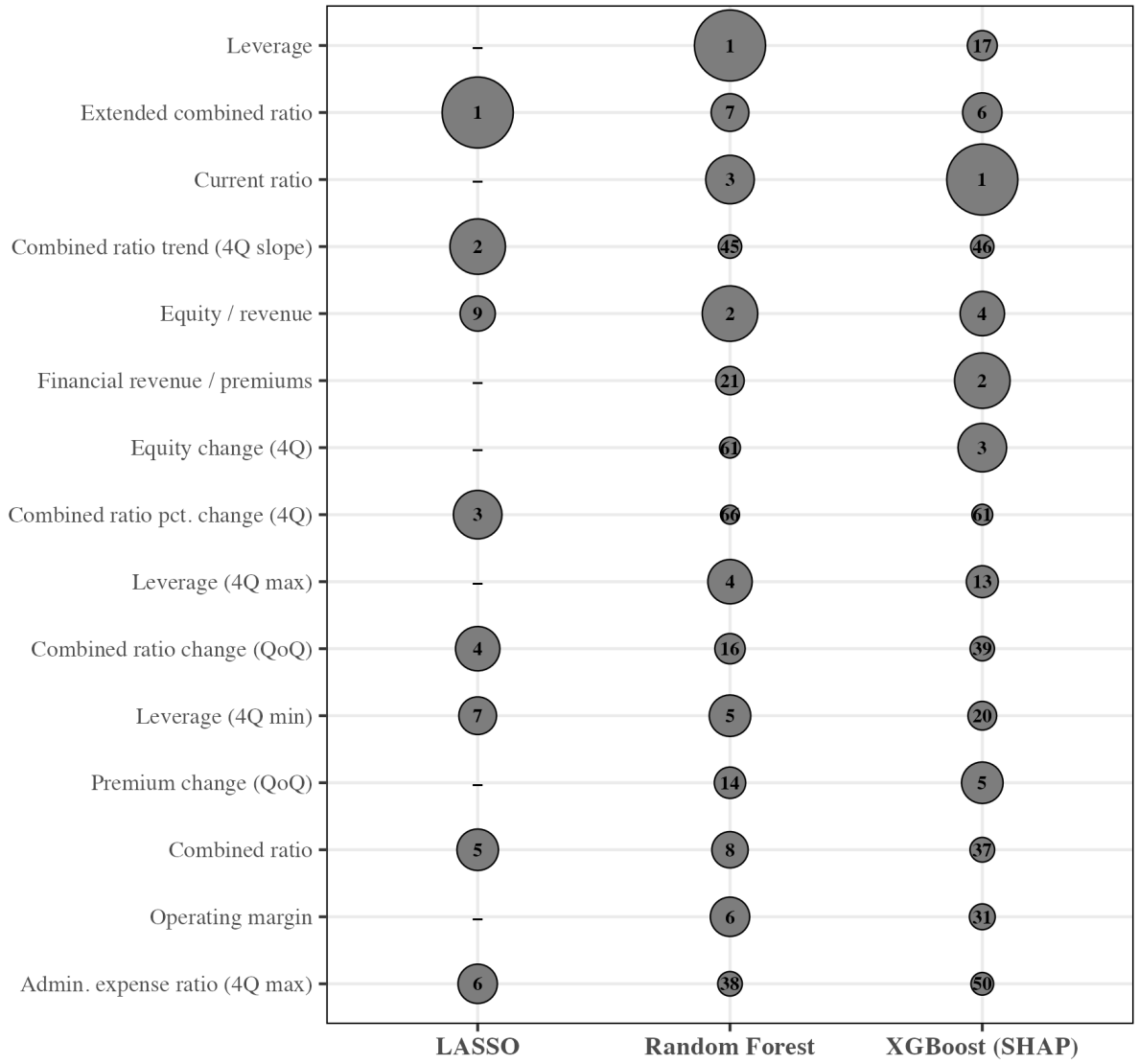


Source: Elaborated by the authors. SHAP values computed from the XGBoost model on the temporal test set (n = 2,000 observations).

Source: Elaborated by the authors.

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

Figure 3. SHAP summary plot for the XGBoost model (top 20 features)

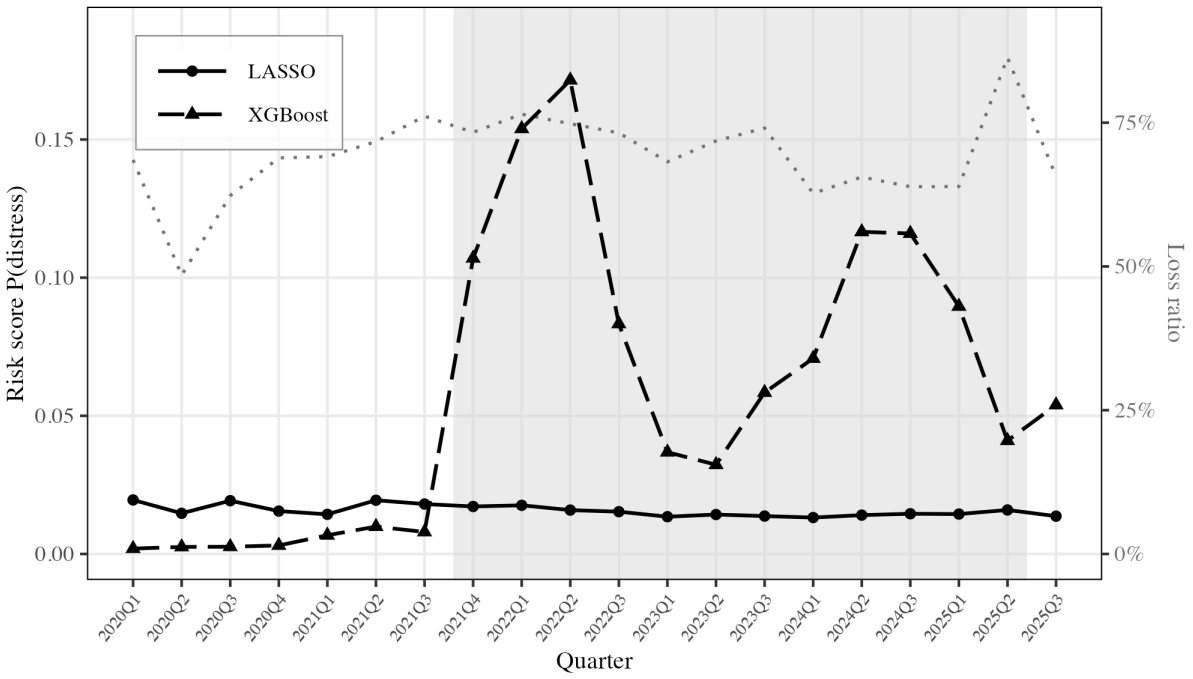


Source: Elaborated by the authors. Numbers indicate feature rank within each model. "–" = feature not selected (LASSO sparsity). Bubble size is inversely proportional to rank.

Source: Elaborated by the authors.

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

Figure 4. Hapvida (ANS 368253) retrospective backtest, 2020Q1–2025Q3

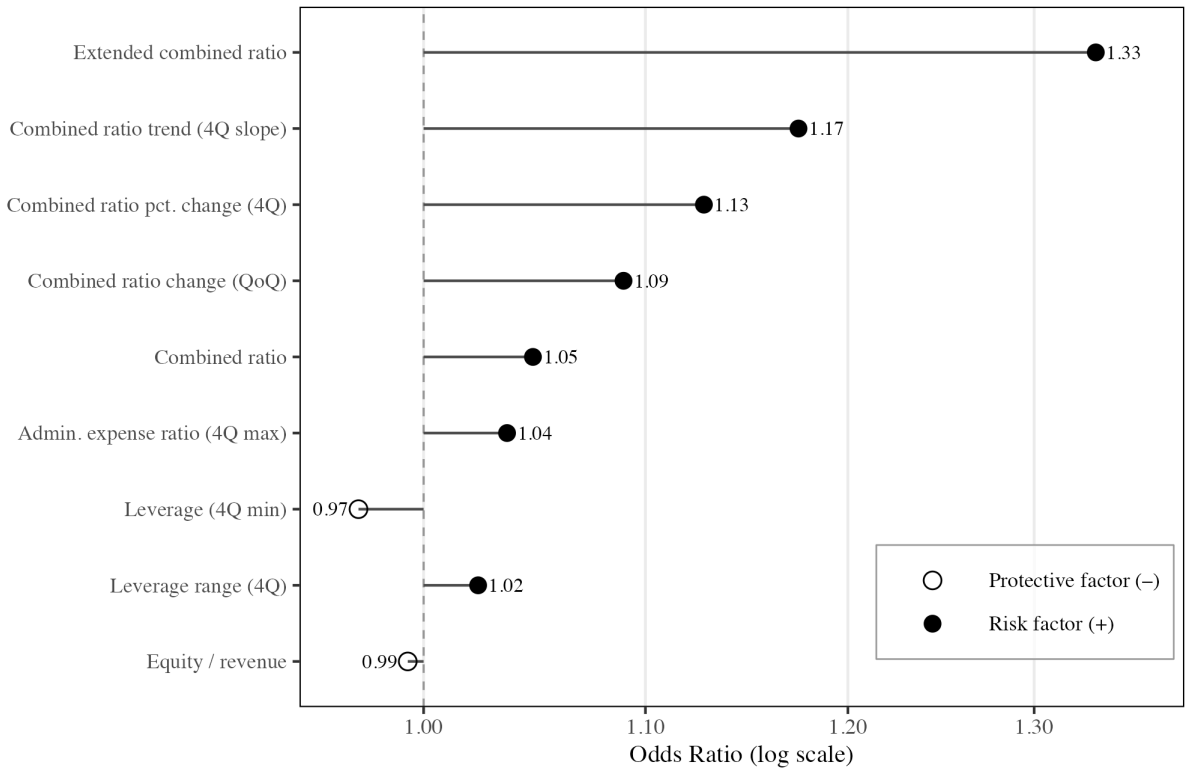


Source: Elaborated by the authors. Shaded area: quarters with combined ratio > 0.95.
Models retrained at each quarter using only prior data (expanding window).
Dotted line: Hapvida's loss ratio (right axis).

Source: Elaborated by the authors.

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.

Figure 5. LASSO logistic regression coefficients for financial distress prediction (odds ratios)



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.

Source: Elaborated by the authors.

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 |coefficient| > 0.005 displayed. Two marginal features (operating margin 1Q change and 4Q slope) omitted.

FIGURE LEGENDS

Figure 1. ROC for the three predictive models on the temporal test set (2023–2025). LASSO logistic regression (dashed line, AUC = 0.716), Random Forest (solid line, AUC = 0.847), and XGBoost (dotted line, AUC = 0.802). The diagonal reference line represents random classification (AUC = 0.5). Shaded areas indicate bootstrap 95% confidence intervals.

Figure 2. LASSO logistic regression coefficient path and selected features. Panel A: regularization path showing coefficient shrinkage as the penalty parameter (λ) increases, with the optimal λ selected by 10-fold cross-validation (vertical dashed line). Panel B: odds ratios with 95% confidence intervals for the 11 features retained at the optimal λ.

Figure 3. SHAP summary plot for the XGBoost model, showing the top 20 features by mean absolute SHAP value. Each point represents one operator-quarter observation; horizontal position indicates the feature's contribution to the predicted log-odds of distress (positive = higher risk); color indicates the feature value (red = high, blue = low).

Figure 4. Hapvida (ANS 368253) retrospective backtest, 2020Q1–2025Q3. Panel A: XGBoost predicted distress probability (solid line, left axis) and percentile rank (bars, right axis). Dashed horizontal line = 75th percentile threshold. Peak signal at 2025Q2 (percentile = 80.7). Panel B: deaccumulated quarterly loss ratio (solid), combined ratio (dashed), and operating margin (dotted). Gray area = COVID-19 pandemic period.

Figure 5. LASSO logistic regression coefficients for financial distress prediction (odds ratios). Eleven features retained by LASSO; nine with |coefficient| > 0.005 shown. Two marginal features (operating margin 1Q change and 4Q slope, both near zero) omitted for clarity. Filled circles denote risk factors (OR > 1); open circles denote protective factors (OR < 1). The extended combined ratio exhibits the strongest effect (OR = 1.33), followed by its four-quarter slope (OR = 1.17) and percentage change (OR = 1.13). Five of nine selected features relate directly to combined ratio dynamics, indicating that cost efficiency trajectory is the most parsimonious linear predictor of operator distress. Leverage (4Q min) and equity-to-revenue ratio are the only protective factors, reflecting capitalization buffers against financial deterioration.

This preprint was submitted under the following conditions:

- The authors declare that the necessary Terms of Free and Informed Consent of participants or patients in the research were obtained and are described in the manuscript, when applicable.
- The authors declare that the preparation of the manuscript followed the ethical norms of scientific communication.
- The authors declare that they are aware that they are solely responsible for the content of the preprint and that the deposit in SciELO Preprints does not mean any commitment on the part of SciELO, except its preservation and dissemination.
- The authors declare that the data, applications, and other content underlying the manuscript are referenced.
- The deposited manuscript is in PDF format.
- The authors declare that the research that originated the manuscript followed good ethical practices and that the necessary approvals from research ethics committees, when applicable, are described in the manuscript.
- The authors declare that once a manuscript is posted on the SciELO Preprints server, it can only be taken down on request to the SciELO Preprints server Editorial Secretariat, who will post a retraction notice in its place.
- The authors agree that the approved manuscript will be made available under a Creative Commons CC-BY license.
- The submitting author declares that the contributions of all authors and conflict of interest statement are included explicitly and in specific sections of the manuscript.
- The authors declare that the manuscript was not deposited and/or previously made available on another preprint server or published by a journal.
- If the manuscript is being reviewed or being prepared for publishing but not yet published by a journal, the authors declare that they have received authorization from the journal to make this deposit.
- The submitting author declares that all authors of the manuscript agree with the submission to SciELO Preprints.