UFPR - Universidade Federal do Paraná

Setor de Ciências Exatas
Departamento de Estatística

Kiron Yago - GRR 20207120

Predição de Insolvência de Empresas Brasileiras de Capital Aberto a partir de Indicadores Contábeis: Uma Abordagem com Modelos Estatísticos e de Aprendizado de Máquina

Curitiba

2025

Sumário

1 Introdução **3**

2 Revisão de Literatura **4**

2.1 Dificuldades Financeiras e Tipos de Crise 4

2.2 Análise de Indicadores Econômico-financeiros 5

2.3 Regressão Logística 6

2.4 LASSO 8

2.5 XGBoost 9

3 Material e Métodos **11**

3.1 Material 11

3.1.1 Conjunto de Dados 11

3.1.2 Recursos Computacionais 11

3.2 Métodos 12

3.2.1 Delineamento da Pesquisa 12

3.2.2 Composição e Seleção das Variáveis 12

4 Resultados e Discussão **14**

4.1 Estatísticas Descritivas 14

4.2 Análise do Modelo Logístico 16

4.3 Análise da Regressão LASSO 17

4.4 Análise do XGBoost 18

5 Considerações Finais **21**

Referências **22**

1 Introdução

O ambiente empresarial brasileiro é caracterizado por intensas oscilações econômicas, instabilidade regulatória e elevados encargos tributários. Nesse cenário, muitas empresas enfrentam dificuldades em manter sua saúde financeira, o que pode levá-las à insolvência e, consequentemente, a processos de Recuperação Judicial (RJ). A Lei nº 11.101/2005, que regula a recuperação judicial e falências no Brasil, foi instituída com o objetivo de preservar empresas viáveis economicamente, permitindo-lhes reestruturar suas dívidas e manter suas atividades produtivas BRASIL (2005).

A identificação precoce de sinais de deterioração financeira é fundamental para a tomada de decisões por investidores, gestores, órgãos reguladores e instituições financeiras. Diante disso, o uso de modelos estatísticos e de aprendizado de máquina para prever o risco de insolvência tem se tornado uma ferramenta relevante na gestão de risco e no suporte à concessão de crédito.

Ao longo das últimas décadas, diversas abordagens foram propostas para modelar o risco de falência. Os estudos pioneiros de Altman (1968), Ohlson (1980) e Zmijewski (1984) introduziram o uso de regressões com base em indicadores financeiros extraídos dos demonstrativos contábeis. Com o avanço da computação, técnicas mais sofisticadas, como regressão penalizada — por exemplo, o método LASSO — e algoritmos de aprendizado de máquina, como o XGBoost, passaram a ser utilizados, permitindo maior capacidade de generalização e seleção automática de variáveis.

Este trabalho tem como objetivo principal desenvolver modelos preditivos capazes de identificar, com base em dados contábeis públicos de empresas brasileiras, a probabilidade de ocorrência de um evento de Recuperação Judicial. Para isso, serão utilizados dados da Comissão de Valores Mobiliários (CVM) entre os anos de 2010 a 2024, e serão aplicadas técnicas de regressão logística, regressão penalizada com LASSO e o algoritmo XGBoost, com ênfase na comparação de desempenho e interpretação dos resultados.

A motivação deste estudo reside na relevância prática de um sistema automatizado de alerta para insolvência, tanto no setor financeiro quanto em órgãos reguladores e ambientes acadêmicos. Espera-se, assim, contribuir com a literatura nacional ao aplicar técnicas modernas de predição sobre uma base extensa e atualizada de dados do mercado brasileiro.

2 Revisão de Literatura

2.1 Dificuldades Financeiras e Tipos de Crise

As dificuldades financeiras enfrentadas por empresas de capital aberto representam um risco relevante para o mercado e para a economia como um todo. Esses episódios podem decorrer de diversos fatores, como má gestão, aumento de endividamento, retração econômica ou variações abruptas nos custos de insumos e financiamentos. O monitoramento e a identificação antecipada desses sinais são, portanto, de interesse tanto de investidores quanto de reguladores e credores.

A Recuperação Judicial (RJ), prevista na Lei nº 11.101/2005, é um dos principais instrumentos legais disponíveis no Brasil para empresas em crise financeira. Esse mecanismo permite que uma empresa renegocie suas dívidas com credores sob supervisão judicial, com o objetivo de preservar a atividade econômica e os empregos, evitando a falência BRASIL (2005). De acordo com a referida legislação, a empresa requerente deve comprovar sua viabilidade econômica e apresentar um plano de reestruturação aprovado pelos credores.

A identificação de empresas com alta probabilidade de solicitar RJ é um desafio preditivo recorrente na literatura de finanças e estatística aplicada. Estudos clássicos sobre falência, como os de Altman (1968), Ohlson (1980) e Zmijewski (1984), utilizam modelos estatísticos baseados em demonstrações contábeis para estimar a probabilidade de insolvência. Essas abordagens têm sido gradualmente complementadas por métodos de aprendizado de máquina que aumentam a capacidade de previsão e reduzem a dependência de pressupostos lineares.

Indicadores como a liquidez corrente, capital de giro sobre ativos, endividamento de curto prazo e retorno sobre o capital investido são amplamente utilizados na modelagem do risco de falência. Estudos recentes, como Carmona et al. (2019), demonstram que algoritmos como o Gradient Boosting apresentam bom desempenho preditivo quando aplicados a séries históricas de dados contábeis. No Brasil, o trabalho de Fuhr (2022) destaca a viabilidade de uso dos dados públicos da CVM para prever eventos de RJ utilizando técnicas modernas de classificação supervisionada.

Além da capacidade preditiva, tais modelos também contribuem para a interpretação dos fatores de risco mais relevantes em contextos de deterioração financeira, o que pode auxiliar instituições financeiras, agências reguladoras e analistas de crédito na formulação de políticas e decisões estratégicas.

2.2 Análise de Indicadores Econômico-financeiros

A análise por meio de indicadores econômico-financeiros é uma prática amplamente difundida nas áreas de contabilidade e finanças, sendo essencial para diagnosticar a saúde financeira de uma empresa e apoiar decisões estratégicas. Esses indicadores, extraídos das demonstrações contábeis, fornecem uma visão objetiva sobre a estrutura de capital, a liquidez, a rentabilidade e a eficiência operacional de uma organização.

Os principais indicadores utilizados em modelos preditivos de falência e insolvência estão tradicionalmente agrupados em quatro categorias: (i) **liquidez**, que mede a capacidade de pagamento no curto prazo; (ii) **endividamento**, que avalia o grau de alavancagem financeira; (iii) **rentabilidade**, que mensura o retorno gerado sobre ativos ou patrimônio; e (iv) **eficiência operacional**, associada ao giro de ativos ou margens de lucro Altman (1968 e Ohlson (1980).

O indicador de **liquidez corrente** (LC), por exemplo, é calculado pela razão entre o ativo circulante e o passivo circulante da empresa. Valores baixos indicam maior risco de inadimplência de curto prazo. Por outro lado, indicadores como **endividamento de curto prazo sobre o total do passivo** (DivCP), ou a razão entre **passivo circulante e passivo total**, indicam o perfil de maturidade das dívidas e sua concentração no curto prazo.

Na dimensão de rentabilidade, destaca-se o **EBIT sobre o total de ativos** (EBIT/Ativo), que avalia o desempenho operacional da empresa em relação aos seus recursos. Já o **retorno sobre o capital próprio** (ROIC ou ROE) é frequentemente utilizado para aferir a eficiência da gestão em gerar lucro para os acionistas.

Além disso, relações compostas como a razão entre a **dívida bruta e o EBIT** (DivBr/EBIT) são interpretadas como proxies do grau de suficiência de caixa operacional para cobrir obrigações financeiras.

A literatura demonstra que a inclusão combinada de diferentes grupos de indicadores amplia a capacidade explicativa dos modelos preditivos Carmona et al. (2019). No presente trabalho, os indicadores foram extraídos diretamente dos demonstrativos padronizados da CVM, e selecionados com base na literatura e em critérios estatísticos de relevância e estabilidade.

2.3 Regressão Logística

A regressão logística é um dos métodos estatísticos mais utilizados na modelagem de variáveis binárias, como no caso da predição de eventos de falência ou recuperação judicial. Sua principal vantagem reside na capacidade de estimar a probabilidade de ocorrência de um evento a partir de um conjunto de variáveis explicativas contínuas ou categóricas, sem requerer a suposição de normalidade dos resíduos Hosmer et al. (2013).

O modelo logístico baseia-se na transformação logística da função linear dos preditores, mapeando os valores estimados para o intervalo (0,1), o que permite a interpretação direta como probabilidades. A equação geral do modelo pode ser expressa como:

$$P(Y = 1 \mid X) = \frac{1}{1 + \exp^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}}$$

Em que Y representa a variável resposta binária (ex: entrada em recuperação judicial), X_i são os preditores (indicadores econômico-financeiros), e β_i os coeficientes estimados. A estimação é feita por máxima verossimilhança, e os coeficientes podem ser interpretados como os efeitos marginais log-odds da variável correspondente.

No contexto de risco de crédito e solvência, a regressão logística tem se mostrado robusta e eficiente, sendo amplamente utilizada por instituições financeiras e reguladores. Trabalhos como o de Ohlson Ohlson (1980) aplicaram o modelo a dados contábeis com grande sucesso preditivo, e sua aplicação permanece relevante, mesmo frente ao avanço de métodos mais complexos como redes neurais e algoritmos de árvore.

Além de sua simplicidade interpretativa, a regressão logística permite a incorporação de técnicas de regularização (como o LASSO) e de validação cruzada para controle de sobreajuste e seleção de variáveis Menard (2002). Também é possível calcular medidas de desempenho como acurácia, sensibilidade, especificidade, F1-score e AUC (Area Under the Curve), o que facilita a comparação com modelos mais sofisticados.

Neste trabalho, a regressão logística é utilizada como modelo de referência para a comparação com métodos mais avançados de classificação, sendo estimada a partir de amostras balanceadas entre eventos e não eventos. Foram realizadas múltiplas simulações bootstrap para avaliação da significância dos coeficientes e da estabilidade das variáveis selecionadas.

A escolha da regressão logística como modelo de referência neste estudo justifica-se por sua ampla aceitação na literatura e na prática profissional na área de risco de crédito e insolvência, além de sua facilidade de implementação, interpretação dos coeficientes e estabilidade em amostras com dimensões moderadas. Trata-se de uma técnica consolidada, o que permite comparar os resultados obtidos com estudos prévios e estabelecer uma base confiável para avaliação de modelos mais complexos. No entanto, é importante reconhecer algumas de suas limitações: a regressão logística pressupõe uma relação linear entre os preditores e o log-odds da variável resposta, o que pode não capturar adequadamente interações ou efeitos não lineares presentes nos dados. Além disso, sua performance pode ser sensível à multicolinearidade entre variáveis e à presença de outliers. Por essas razões, outros métodos mais flexíveis, como o LASSO e o XGBoost, são explorados neste

trabalho de forma complementar, com o objetivo de avaliar se oferecem ganhos de desempenho e interpretabilidade no contexto da predição de insolvência empresarial.

2.4 LASSO

O método LASSO (*Least Absolute Shrinkage and Selection Operator*) foi proposto por Tibshirani Tibshirani (1996) como uma extensão da regressão linear tradicional, com o objetivo de realizar simultaneamente **regularização** e **seleção de variáveis**. No contexto da regressão logística, sua aplicação é particularmente útil quando há um grande número de variáveis explicativas e a necessidade de evitar sobreajuste (*overfitting*).

A formulação do LASSO inclui um termo de penalização que impõe uma restrição à soma dos valores absolutos dos coeficientes do modelo. Essa penalização força alguns coeficientes a se tornarem exatamente zero, promovendo uma forma automática de seleção de variáveis. No caso da regressão logística penalizada, a função de custo é dada por:

$$\min \left\{ -\ell(\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

onde $\ell(\beta)$ é a log-verossimilhança da regressão logística, λ é o parâmetro de penalização e β_j são os coeficientes a serem estimados. O parâmetro λ controla a intensidade da penalização: valores maiores promovem modelos mais parcimoniosos, com menos variáveis selecionadas Friedman et al. (2010).

A principal vantagem do LASSO está na sua capacidade de lidar com **colinearidade entre variáveis** e de proporcionar modelos mais interpretáveis. Em aplicações com dados financeiros, isso é particularmente relevante, dado que muitos indicadores contábeis são altamente correlacionados.

Em problemas de classificação binária, como o estudo da insolvência ou recuperação judicial de empresas, o LASSO permite a construção de modelos robustos com base em subconjuntos relevantes de indicadores, eliminando automaticamente os menos informativos Fan; Li (2001). Além disso, sua implementação computacional é eficiente e amplamente disponível em pacotes estatísticos como `glmnet` no R.

Neste trabalho, o LASSO é utilizado como uma etapa de seleção de variáveis, complementando a análise com regressão logística clássica. Foram realizadas simulações em larga escala com diferentes amostras balanceadas para identificar os preditores com maior frequência de seleção e avaliar sua estabilidade estatística.

2.5 XGBoost

O algoritmo **XGBoost** (*Extreme Gradient Boosting*), proposto por Chen e Guestrin Chen; Guestrin (2016), é uma das técnicas mais populares de aprendizado de máquina supervisionado para tarefas de classificação e regressão. Ele pertence à família dos modelos de **boosting** baseados em árvores de decisão, sendo conhecido por sua alta acurácia, velocidade e capacidade de lidar com dados estruturados.

O XGBoost constrói um conjunto de árvores de decisão de forma sequencial, onde cada nova árvore é ajustada para corrigir os erros residuais cometidos pelas árvores anteriores. A função objetivo do modelo é composta por dois termos: a **função de perda** (geralmente log-loss ou erro quadrático) e um **termo de regularização** que penaliza a complexidade do modelo:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k)$$

em que l é a função de perda, Ω é a penalização por complexidade, e f_k representa cada árvore do modelo.

Entre as principais **vantagens** do XGBoost estão:

- Suporte nativo a **valores ausentes**;
- Capacidade de capturar **interações não lineares** entre variáveis;
- Regularização por penalização da profundidade e do número de folhas da árvore;
- Interpretação da **importância das variáveis** por meio de métricas como `Gain`, `Cover` e `Frequency`.

Em contextos financeiros, especialmente na predição de insolvência e recuperação judicial, o XGBoost tem se mostrado superior a modelos lineares tradicionais em termos de capacidade preditiva, conforme apontado por estudos como Carmona et al. (2019) e Fuhr (2022). Seu uso permite capturar relações complexas e interações entre indicadores contábeis, proporcionando previsões mais precisas.

No presente trabalho, o XGBoost foi implementado com validação cruzada e avaliação por métricas como AUC, sensibilidade, especificidade e F1-score, possibilitando comparação direta com os modelos de regressão logística.

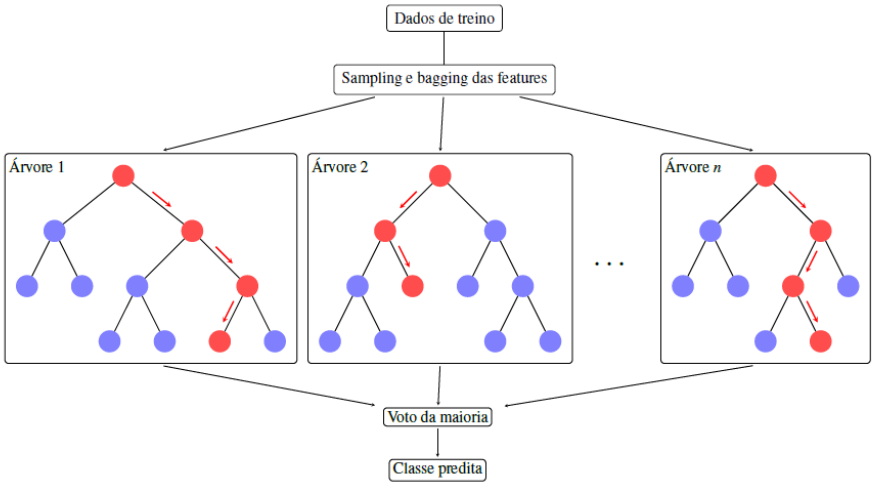

Figura 1: Ilustração do funcionamento sequencial do Boosting

A figura acima ilustra o funcionamento de um modelo baseado em conjunto de árvores de decisão, como o Random Forest. O processo inicia-se com os dados de treino originais, a partir dos quais são geradas diversas amostras por meio de técnicas de **bootstrap** (re-amostragem com reposição) e **bagging de variáveis** (seleção aleatória de subconjuntos de features para cada árvore).

Cada uma dessas amostras é utilizada para treinar uma árvore de decisão distinta, formando um conjunto de modelos individuais (Árvore 1, Árvore 2, …, Árvore *n*). Esse processo contribui para a redução da variância do modelo, uma vez que árvores treinadas em subconjuntos distintos dos dados tendem a cometer erros diferentes entre si.

Durante a fase de predição, as saídas de todas as árvores são combinadas por meio de **votação da maioria** (para problemas de classificação), resultando na classe predita final. Esse mecanismo de agregação permite que o ensemble tenha melhor desempenho preditivo e maior estabilidade em relação a uma única árvore, sendo menos suscetível a overfitting. As setas vermelhas nas árvores indicam os caminhos de decisão tomados internamente com base nas variáveis selecionadas em cada nó.

Esse tipo de abordagem se destaca por sua robustez e capacidade de lidar com dados complexos, com múltiplas variáveis e possíveis interações não lineares entre elas.

3 Material e Métodos

3.1 Material

3.1.1 Conjunto de Dados

O conjunto de dados utilizado neste trabalho foi construído a partir das demonstrações financeiras padronizadas de companhias abertas brasileiras, disponibilizadas pela Comissão de Valores Mobiliários (CVM). As informações contemplam o período de 2010 a 2024, abrangendo empresas de diferentes setores econômicos e portes.

Foram utilizadas as seguintes demonstrações contábeis: Balanço Patrimonial (ativo e passivo), Demonstração do Resultado do Exercício (DRE), Demonstração de Fluxo de Caixa (DFC) e Demonstração das Mutações do Patrimônio Líquido (DMPL). A partir dessas, foram derivados indicadores financeiro-contábeis relacionados à liquidez, endividamento, rentabilidade e estrutura de capital, conforme práticas sugeridas por Altman Altman (1968) e Ohlson Ohlson (1980).

A variável resposta (`rj`) foi construída com base na identificação de empresas que entraram em Recuperação Judicial (RJ) entre os anos analisados. A classificação foi feita por meio de busca textual nos nomes das companhias, com auxílio de web scraping de fontes públicas (ex. JusBrasil), e posterior validação manual.

Após tratamento e limpeza dos dados, a base final foi composta por 5.933 observações, referentes a 693 empresas listadas na CVM.

3.1.2 Recursos Computacionais

Todas as análises foram desenvolvidas na linguagem R (versão 4.x), utilizando pacotes como `tidyverse`, `glmnet`, `xgboost`, `broom`, `yardstick`, `modelr`, entre outros. O tratamento e integração dos dados foi realizado com uso de pipelines organizados, assegurando reprodutibilidade dos resultados. Scripts automatizados foram utilizados para a coleta de dados, transformação de variáveis e aplicação dos modelos.

3.2 Métodos

3.2.1 Delineamento da Pesquisa

Este trabalho classifica-se como uma pesquisa aplicada e quantitativa, com abordagem empírica e foco preditivo. Seu objetivo principal é investigar a capacidade de modelos estatísticos e de aprendizado de máquina em prever a ocorrência de Recuperação Judicial (RJ) de empresas com base em indicadores extraídos de demonstrações contábeis.

O delineamento metodológico adotado foi orientado por uma estratégia de reamostragem. Considerando que apenas 24 observações no conjunto de dados correspondem a empresas que passaram por RJ, há um forte desbalanceamento da variável resposta. Para lidar com isso, cada modelo foi ajustado sobre 1.000 amostras balanceadas obtidas por bootstrap estratificado — ou seja, a cada simulação, todas as observações com `rj = 1` foram combinadas com um número igual de controles (`rj = 0`) selecionados aleatoriamente sem reposição.

Três abordagens principais foram utilizadas:

- **Regressão Logística**: utilizada como modelo base, pela sua interpretabilidade e capacidade de estimar probabilidades associadas à ocorrência de RJ;
- **Regressão LASSO**: incorporada como método de seleção de variáveis, com penalização L1 para reduzir multicolinearidade e identificar preditores relevantes;
- **XGBoost**: modelo baseado em árvores de decisão em boosting sequencial, capaz de capturar interações complexas e não linearidades nos dados.

As simulações permitiram estimar a robustez dos modelos em diferentes amostras, aferir a estabilidade dos coeficientes, calcular a frequência de significância estatística e avaliar o desempenho por meio de métricas como Acurácia, Precisão, F1-score, Sensibilidade, Especificidade e AUC.

3.2.2 Composição e Seleção das Variáveis

As variáveis explicativas utilizadas foram construídas a partir dos demonstrativos contábeis das empresas. Com base na literatura especializada em predição de insolvência Altman (1968); Ohlson (1980); Zmijewski (1984), foram inicialmente considerados 11 indicadores financeiros:

Tabela 1: Descrição das variáveis explicativas utilizadas no modelo.

Variável	Fórmula	Descrição
DivCP	Passivo Circulante / Passivo Total	Proporção da dívida de curto prazo sobre o total de obrigações.
LC	Ativo Circulante / Passivo Circulante	Indica a capacidade de liquidez de curto prazo.
FCO_At	Fluxo de Caixa Operacional / Ativo Total	Geração de caixa operacional em relação ao tamanho da empresa.
DivBr_At	Dívida Bruta / Ativo Total	Nível de endividamento total em relação ao ativo.
PL_DivBr	Patrimônio Líquido / Dívida Bruta	Capacidade de cobertura da dívida com recursos próprios.
CapGir_At	(Ativo Circulante - Passivo Circulante) / Ativo Total	Proporção do capital de giro em relação ao ativo.
GiroAt	Receita Líquida / Ativo Total	Eficiência da empresa em gerar receita com seus ativos.
EBIT_At	EBIT / Ativo Total	Rentabilidade operacional sobre o total de ativos.
DivBr_Ebit	Dívida Bruta / EBIT	Grau de comprometimento da dívida sobre o resultado operacional.
ROICf	EBIT / Patrimônio Líquido	Retorno sobre o capital próprio.
LN_At	log(Ativo Total)	Tamanho da empresa em escala logarítmica.

A seleção de variáveis para os modelos finais seguiu dois critérios quantitativos:

1. **Regressão Logística por Reamostragem**: cada variável foi avaliada quanto à frequência com que seu coeficiente foi estatisticamente significativo ($p < 0.05$) ao longo das 1.000 simulações;
2. **Seleção via LASSO**: registrou-se o percentual de vezes em que a variável foi retida no modelo ajustado com penalização, indicando sua relevância na presença de regularização.

As variáveis mais frequentes em ambos os métodos foram mantidas como preditores no modelo XGBoost. Isso garantiu uma seleção estável e parcimoniosa, priorizando indicadores robustos à variação amostral e relevantes do ponto de vista econômico-financeiro.

4 Resultados e Discussão

4.1 Estatísticas Descritivas

A análise descritiva das variáveis explicativas é uma etapa fundamental para compreender a natureza dos dados utilizados nos modelos preditivos. No entanto, a base empregada neste estudo apresenta características desafiadoras, como escalas heterogêneas entre variáveis, presença de valores extremos (outliers) e forte assimetria em diversas distribuições.

Para mitigar parte dessas distorções, foi realizada a **padronização das variáveis explicativas**, transformando-as para apresentar média zero e desvio padrão um. Essa transformação é especialmente útil em métodos como **regressão LASSO** e **XGBoost**, que podem ser sensíveis a escalas distintas.

As Figura abaixo apresenta a distribuição das variáveis explicativas em boxplots.

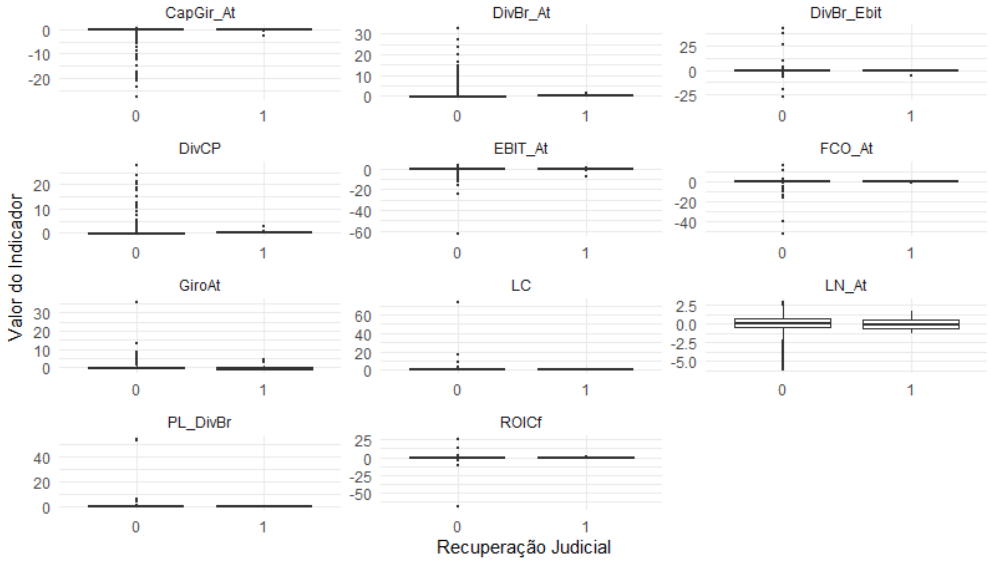

Figura 2: Distribuição dos indicadores econômico-financeiros por situação de Recuperação Judicial

A Figura 2 apresenta a distribuição dos indicadores econômico-financeiros considerados no estudo, segmentados entre empresas que entraram em Recuperação Judicial (rj = 1) e aquelas que permaneceram solventes (rj = 0). Os gráficos do tipo boxplot evidenciam um comportamento altamente heterogêneo entre os indicadores, marcado por forte presença de outliers e assimetrias.

Observa-se que diversas variáveis possuem caudas longas ou concentrações assimétricas, o que resulta em boxplots comprimidos e com pouco contraste visual entre os grupos. Isso dificulta a identificação clara de padrões discriminantes entre empresas solventes e em Recuperação Judicial apenas com base em análise exploratória.

A presença de valores extremos, como nas variáveis `EBIT_At`, `DivBr_At` e `ROICf`, destaca a diversidade das estruturas financeiras das empresas brasileiras, refletindo a heterogeneidade do ambiente empresarial analisado. Em contrapartida, algumas variáveis, como `LN_At` e `LC`, apresentam distribuições mais concentradas e simétricas, sugerindo maior estabilidade estatística.

Diante dessa complexidade, torna-se evidente a necessidade do uso de modelos estatísticos mais robustos, capazes de lidar com dispersões elevadas, escalas distintas e potenciais relações não lineares. Por esse motivo, o presente trabalho adota a regressão logística como modelo inicial de referência, complementada por técnicas de regularização (LASSO) e algoritmos mais flexíveis, como o XGBoost, para realizar a seleção de variáveis e melhorar a capacidade preditiva frente à diversidade dos dados.

4.2 Análise do Modelo Logístico

A regressão logística foi empregada como modelo base para a predição da probabilidade de uma empresa entrar em recuperação judicial. Esse modelo é amplamente utilizado para problemas de classificação binária, sendo interpretável e estatisticamente sólido Hosmer et al. (2013).

A função de verossimilhança da regressão logística é dada por:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

onde p_i representa a probabilidade da i-ésima empresa entrar em recuperação judicial (rj = 1), e os X_{ij} são os valores das variáveis explicativas selecionadas.

Neste estudo, o modelo foi avaliado por meio de 5.000 amostragens balanceadas da base `var_econfin`, com igual número de empresas que entraram e não entraram em recuperação judicial. Em cada simulação, ajustou-se o modelo logístico e armazenaram-se os coeficientes e valores-p.

A Tabela 1 apresenta os coeficientes médios estimados para cada variável explicativa, juntamente com a proporção de vezes em que os coeficientes foram estatisticamente significativos ao nível de 5%. Essa abordagem permite verificar a **robustez estatística** das variáveis e sua relevância na explicação do desfecho.

Tabela 2: Média dos coeficientes estimados e proporção de significância para cada variável no modelo logístico.

Variável	Coef. Médio	Prop. Significância
EBIT_At	-1.846148e+14	0.309
CapGir_At	7.198545e+14	0.229
DivCP	5.565114e+14	0.229
LC	-1.399297e+15	0.227
LN_At	1.009241e+13	0.196
(Intercept)	6.003748e+12	0.185
DivBr_At	1.777605e+14	0.183
FCO_At	7.789850e+13	0.180
PL_DivBr	1.162532e+15	0.179
GiroAt	-6.969505e+13	0.169
DivBr_Ebit	-8.649766e+13	0.169
ROICf	-8.949214e+12	0.167

A Tabela 3 apresenta os resultados consolidados de 5.000 simulações da regressão logística com amostras balanceadas. Para cada variável explicativa, são apresentados o **coeficiente médio estimado** e a **proporção de simulações em que o coeficiente foi estatisticamente significativo ao nível de 5%**.

O objetivo principal da regressão logística neste estudo não foi a previsão em si, mas sim a **avaliação da robustez estatística das variáveis**, com foco em sua utilização como **etapa de seleção preliminar de atributos**.

Nota-se que a variável `EBIT_At` (EBIT sobre Ativo Total) obteve a maior proporção de significância (30,9%), com seu coeficiente de sinal negativo, o que está alinhado com a literatura: **empresas com menor rentabilidade operacional tendem a ter maior risco de insolvência** Altman (1968 e Ohlson (1980).

Outras variáveis que se destacam pela frequência de significância acima de 20% incluem: - `CapGir_At`: capital de giro em relação ao ativo total; - `DivCP`: dívida de curto prazo sobre o passivo total; - `LC`: liquidez corrente; - `LN_At`: logaritmo do ativo total (proxy de tamanho).

Esses resultados reforçam a **importância de indicadores de liquidez, estrutura de capital e rentabilidade** no contexto de risco de recuperação judicial, servindo como base sólida para a etapa de modelagem com regularização (LASSO) e métodos não-lineares (XGBoost).

Deve-se destacar que o valor absoluto dos coeficientes estimados aparece inflado (ordem de 10^{13} a 10^{14}), o que pode estar relacionado a **eventuais problemas de multicolinearidade**, muito comuns em estudos desbalanceados. No entanto, como o foco está na **frequência de significância** e não na magnitude dos coeficientes, o uso do modelo logístico continua válido como **ferramenta de triagem e diagnóstico**.

Assim, essa etapa cumpriu seu papel central ao indicar um subconjunto de variáveis com **relevância estatística recorrente**, direcionando os próximos modelos para um conjunto mais parcimonioso e informativo.

4.3 Análise da Regressão LASSO

O modelo LASSO (Least Absolute Shrinkage and Selection Operator), conforme introduzido por Tibshirani Tibshirani (1996), foi empregado neste trabalho com o propósito de auxiliar a **seleção de variáveis** em um contexto de classificação binária, visando prever a entrada de empresas em **recuperação judicial (RJ)**.

A versão penalizada da regressão logística utilizada é definida pela função de custo:

$$\min \left\{ -\ell(\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

onde $\ell(\beta)$ é a log-verossimilhança do modelo, λ é o hiperparâmetro de penalização, e β_j são os coeficientes estimados. A penalização L1 induz **esparsidade nos coeficientes**, o que implica que variáveis com menor contribuição preditiva têm seus coeficientes reduzidos a zero, sendo assim excluídas do modelo.

Neste estudo, foram realizadas 1.000 simulações com amostras balanceadas da base `var_econfin`. Para cada uma, ajustou-se um modelo LASSO com validação cruzada para selecionar o melhor λ e identificou-se o conjunto de variáveis selecionadas.

A Tabela 4 apresenta a proporção de vezes em que cada variável foi selecionada (± coeficiente diferente de zero) ao longo das simulações, refletindo sua **importância preditiva recorrente** no modelo penalizado.

Tabela 3: Frequência de seleção e média dos coeficientes distintos de zero para cada variável no modelo LASSO.

variavel	vezes_selecionada	prop	media_coef
(Intercept)	1000	1.000	-0.157
LC	819	0.819	-0.675
PL_DivBr	779	0.779	-1.226
DivBr_Ebit	512	0.512	-0.003
DivBr_At	328	0.328	0.614
LN_At	319	0.319	0.150
DivCP	263	0.263	1.807
ROICf	261	0.261	-0.024
GiroAt	195	0.195	-0.192
EBIT_At	171	0.171	-5.416
FCO_At	161	0.161	-2.962
CapGir_At	128	0.128	-0.404

A análise da Tabela 4 permite destacar algumas variáveis com **alta estabilidade estatística**, como *LC* e *PL_DivBr*, que foram selecionadas em mais de 75% das iterações. Tais resultados indicam que essas variáveis possuem **forte poder preditivo** e recorrência no processo de ajuste do modelo.

Outras variáveis, como *DivBr_Ebit* e *DivBr_At*, também apresentaram frequências moderadas de seleção, o que justifica sua manutenção na próxima etapa. Por outro lado, variáveis com baixa frequência de seleção, como *FCO_At* e *CapGir_At*, foram consideradas **menos relevantes estatisticamente**.

Com base nesses resultados, foram selecionadas as seguintes variáveis para compor o modelo XGBoost: *LC*, *PL_DivBr*, *DivCP*, *CapGir_At*, *EBIT_At*, *DivBr_Ebit* e *LN_At*

Essa escolha prioriza a parcimônia e robustez do modelo, fundamentando-se na **frequência de seleção** observada no LASSO, ponderando também os importantes resultados que tivemos na Regressão Logística.

4.4 Análise do XGBoost

A construção do modelo considerou as sete variáveis que apresentaram **maior frequência de seleção no LASSO** e, simultaneamente, **alta proporção de significância estatística** nos modelos de regressão logística simulados. A escolha baseou-se na **consistência entre os métodos**, aliada à **relevância econômica dos indicadores**, buscando robustez e interpretabilidade no modelo final.

Após a seleção de variáveis, o modelo XGBoost foi utilizado para prever a probabilidade de uma empresa entrar em recuperação judicial (rj = 1). A escolha do XGBoost se justifica por sua capacidade de lidar com relações não lineares e variáveis com diferentes escalas, além de oferecer alta performance preditiva.

Tabela 4: Importância das variáveis no modelo XGBoost ao longo de 1000 simulações.

Feature	ganho_medio
EBIT_At	0.238
PL_DivBr	0.237
DivBr_Ebit	0.129
DivCP	0.124
LC	0.123
CapGir_At	0.118
LN_At	0.030

Esses resultados reforçam a relevância de indicadores de **rentabilidade**, **estrutura de capital** e **liquidez de curto prazo** para a predição de recuperação judicial, corroborando achados da literatura e evidências obtidas nas análises com regressão logística e LASSO.

- **EBIT sobre Ativo Total** foi a variável com maior ganho médio, sugerindo que empresas com rentabilidade operacional reduzida tendem a estar associadas a maior risco de recuperação judicial. Esse achado é coerente com a literatura, uma vez que lucros operacionais negativos limitam a geração de caixa e o cumprimento de obrigações.
- **Patrimônio Líquido sobre Dívida Bruta** também apresentou alto ganho, refletindo a capacidade de cobertura da dívida com capital próprio, um importante sinal de solvência.
- Variáveis como **Dívida Bruta sobre EBIT** e **Dívida de Curto Prazo sobre o Passivo Total** também se destacaram, indicando a relevância da estrutura de capital na previsão do desfecho.
- Indicadores de **liquidez de curto prazo**, como **Liquidez Corrente** e **Capital de Giro sobre o Ativo Total**, mostraram-se relevantes para a discriminação entre os grupos.
- Por fim, **Logaritmo do Ativo Total**, embora tenha apresentado o menor ganho médio, foi selecionado em 996 das 1.000 simulações, o que justifica sua inclusão por critérios de consistência estatística e controle de escala.

Esses achados reforçam a capacidade do XGBoost em capturar relações complexas e interações entre variáveis, oferecendo um modelo preditivo competitivo e interpretável para o problema de classificação do risco de recuperação judicial de empresas.

A performance do modelo XGBoost foi avaliada em 1.000 simulações balanceadas, gerando as métricas média e intervalos de confiança (IC 95%) para acurácia, precisão, sensibilidade, especificidade, F1 e AUC.

Tabela 5: Desempenho médio do modelo XGBoost

Metrica	media	ic_inf	ic_sup
Acuracia	0.7623125	0.7554052	0.7692198
Precisao	0.7803240	0.7720780	0.7885699
Sensibilidade	0.7710000	0.7609656	0.7810344
Especificidade	0.7536250	0.7411114	0.7661386
F1_Score	0.7616269	0.7545717	0.7686821
AUC	0.8685469	0.8628963	0.8741974

A **acurácia média** foi de aproximadamente **76%**, sugerindo que o modelo consegue classificar corretamente, em média, três a cada quatro observações. A **sensibilidade** também se destacou, com média em torno de **77%**, o que demonstra uma boa capacidade do modelo em identificar corretamente empresas que entraram em recuperação judicial — uma prioridade neste tipo de estudo.

A **especificidade**, por sua vez, foi ligeiramente inferior, com média próxima de **75%**, apontando que o modelo é ligeiramente mais eficaz em detectar empresas em risco do que em confirmar aquelas saudáveis. A **precisão** (cerca de **78%**) confirma essa tendência, indicando que a maioria das empresas classificadas como insolventes de fato pertencem à classe positiva.

A métrica **F1 Score**, que combina precisão e sensibilidade em um único índice harmônico, também apresentou valores consistentes, o que sugere **bom equilíbrio entre falsos positivos e falsos negativos**. Por fim, a **AUC média de 0.869** reflete a capacidade geral do modelo em discriminar corretamente entre os dois grupos, sendo considerada **excelente** segundo os critérios da literatura de classificação binária.

Em conjunto, essas evidências apontam que o modelo XGBoost, ajustado com as variáveis selecionadas a partir da regressão logística e do LASSO, é uma ferramenta eficaz para predição de risco de insolvência, conciliando desempenho preditivo e estabilidade estatística.

5 Considerações Finais

O presente estudo teve como principal objetivo **identificar os indicadores financeiros com maior poder explicativo** para sinalizar a **insolvência de empresas brasileiras de capital aberto**, tendo como evento de referência a entrada em **recuperação judicial**. Para isso, utilizou-se uma base de dados estruturada a partir dos demonstrativos contábeis divulgados à Comissão de Valores Mobiliários (CVM), no período de 2010 a 2024.

Inicialmente, foram aplicadas técnicas de análise descritiva, evidenciando diferenças entre empresas solventes e insolventes em variáveis como **liquidez corrente**, **alavancagem**, **retorno sobre ativos** e **endividamento**. Na sequência, foram utilizados modelos estatísticos para seleção e avaliação da relevância das variáveis explicativas, com destaque para:

- **Regressão logística simulada**, que forneceu insights sobre a robustez e significância estatística dos coeficientes estimados;
- **LASSO**, que permitiu a seleção automática de variáveis com base em sua relevância preditiva e estabilidade;
- **XGBoost**, que ofereceu a análise final de importância relativa das variáveis, por meio do ganho médio de informação.

A análise empírica revelou que os indicadores mais fortemente associados à probabilidade de insolvência foram aqueles relacionados à **rentabilidade sobre ativos**, **nível de endividamento de curto prazo**, e **capacidade de geração operacional**. O modelo XGBoost obteve desempenho satisfatório, com **AUC média de 0.869**, evidenciando sua robustez para o problema de classificação.

Diferentemente de estudos com foco na previsão individual de falência, este trabalho enfatizou a **interpretação dos preditores mais relevantes**, contribuindo para o entendimento dos mecanismos financeiros que precedem situações de crise empresarial.

Como possíveis direções para estudos futuros, destacam-se:

- A realização de análises segmentadas por **setor de atuação**, **porte da empresa** ou **ciclos econômicos**;
- A incorporação de **variáveis qualitativas ou de mercado**, como governança corporativa, indicadores de crédito e rating;
- A aplicação de modelos **longitudinais ou dinâmicos**, que permitam avaliar a evolução temporal do risco de insolvência.

Com isso, espera-se que os resultados aqui obtidos possam servir como base para **diagnósticos financeiros precoces**, subsidiando decisões de crédito, investimentos e estratégias de monitoramento empresarial.

Referências

ALTMAN, E. I. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. **The Journal of Finance**, v. 23, n. 4, p. 589–609, 1968. Wiley.

BRASIL. Lei nº 11.101, de 9 de fevereiro de 2005. Regula a recuperação judicial, a extrajudicial e a falência do empresário e da sociedade empresária., 2005. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2005/lei/l11101.htm>.

CARMONA, C.; EPPRECHT, E.; MATOS, P. Predição de insolvência por meio de algoritmos de aprendizado de máquina. **Revista de Contabilidade e Organizações**, v. 13, n. 36, p. 1–15, 2019.

CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.**Anais...** . p.785–794, 2016. ACM.

FAN, J.; LI, R. Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. **Journal of the American Statistical Association**, v. 96, n. 456, p. 1348–1360, 2001.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. **Journal of Statistical Software**, v. 33, n. 1, p. 1–22, 2010.

FUHR, G. T. Predição de empresas em Recuperação Judicial por meio de demonstrativos financeiros públicos utilizando aprendizado de máquina., 2022. Trabalho de Conclusão de Curso - Universidade Federal do Rio Grande do Sul.

HOSMER, D. W.; LEMESHOW, S.; STURDIVANT, R. X. **Applied Logistic Regression**. 3º ed. John Wiley & Sons, 2013.

MENARD, S. **Applied Logistic Regression Analysis**. SAGE Publications, 2002.

OHLSON, J. A. Financial Ratios and the Probabilistic Prediction of Bankruptcy. **Journal of Accounting Research**, v. 18, n. 1, p. 109–131, 1980. Wiley.

TIBSHIRANI, R. Regression Shrinkage and Selection via the Lasso. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 58, n. 1, p. 267–288, 1996.

ZMIJEWSKI, M. E. Methodological Issues Related to the Estimation of Financial Distress Prediction Models. **Journal of Accounting Research**, v. 22, p. 59–82, 1984.