ISSN 1808-057X
DOI: 10.1590/1808-057x202010360

ARTIGO ORIGINAL

Avaliação de falências de empresas por meio de florestas causais

Wanderson Rocha Bittencourt[1]
https://orcid.org/0000-0003-3417-2225
E-mail: wandersonrochab@yahoo.com.br

Pedro H. M. Albuquerque[1]
https://orcid.org/0000-0002-1415-716X
E-mail: pedroa@unb.br

[1] Universidade de Brasília, Faculdade de Administração, Contabilidade, Economia e Gestão de Políticas Públicas, Departamento de Administração, Brasília, DF, Brasil

Recebido em 06.08.2019 – *Desk* aceite em 09.09.2019 – 3ª versão aprovada em 29.01.2020 – *Ahead of print* em 10.07.2020
Editora Associada: Fernanda Finotti Cordeiro Perobelli

RESUMO

Esta pesquisa buscou analisar as variáveis que podem influenciar a falência das empresas. Durante vários anos, as principais pesquisas sobre falência reportaram as metodologias convencionais visando à sua predição. Em suas análises, a utilização de variáveis contábeis predominou maciçamente. Porém, ao aplicá-las, as variáveis contábeis eram consideradas homogêneas, ou seja, para os modelos tradicionais, presumia-se que em todas as empresas o comportamento dos indicadores era similar, ignorando a heterogeneidade entre elas. Observa-se, ainda, a relevância da crise financeira ocorrida no final de 2007, causando grande colapso financeiro mundial, tendo efeitos diferentes nos mais diversos setores e empresas. Nesse cenário, pesquisas que visam identificar problemas como a heterogeneidade entre as empresas e analisar as diversidades entre elas ganham relevância, haja vista que as características setoriais de estrutura de capital, porte, dentre outras, variam de acordo com as empresas. A partir disso, novas abordagens aplicadas à modelagem de previsão de falência devem considerar a heterogeneidade entre as empresas, buscando aprimorar ainda mais as modelagens utilizadas. Foram utilizadas a árvore e a floresta causais com dados contábeis trimestrais e setoriais de 1.247 empresas, sendo 66 falidas, das quais 44 depois de 2008 e 22 antes. Os resultados mostraram que existe heterogeneidade não observada quando se analisam os processos de falência das empresas, colocando em cheque os modelos tradicionais como, por exemplo, análise discriminante e *logit*, dentre outros. Por conseguinte, com o elevado volume em dimensões, observou-se que pode haver uma forma funcional capaz de explicar a falência das empresas, porém essa não é linear. Destaca-se, ainda, que existem setores mais propensos a crises financeiras, agravando o processo de falência.

Palavras-chave: floresta causal, árvore causal, heterogeneidade, crise financeira, falência.

Endereço para correspondência

Wanderson Rocha Bittencourt
Universidade de Brasília, Faculdade de Administração, Contabilidade, Economia e Gestão de Políticas Públicas, Departamento de Ciências Contábeis e Atuariais
Campus Universitário Darcy Ribeiro, Bloco A-2 – CEP 70910-900
Asa Norte – Brasília – DF – Brasil

1. INTRODUÇÃO

Por várias vezes, as pesquisas empíricas têm como foco a estrutura, a causalidade ou o tratamento de um fenômeno de interesse. Na economia, por exemplo, algumas pesquisas buscam analisar os efeitos de uma política econômica no desenvolvimento econômico e na empregabilidade, dentre outros. Contudo, existem condições não observáveis que inviabilizam a estratégia, obtendo efeitos indesejados (Belloni, Chernozhukov & Hansen, 2014a).

Com esse cenário, os recursos computacionais ganham espaço, sendo inevitável sua aplicação em contextos como a economia e as finanças. Os sistemas computacionais estão auxiliando a análise de grandes bancos de dados (*big data*) em que as ferramentas estatísticas convencionais, como a análise de regressão, apresentam resultados aquém de outras ferramentas (Varian, 2014, 2016).

Com as ferramentas estatísticas tradicionais (regressões), a manipulação dos dados e o posterior potencial de previsão ficam restritos, principalmente, a modelos lineares, não capturando as relações com outros comportamentos. Nessa mesma linha de pensamento, as pesquisas empíricas geralmente reportam suas estimativas baseadas em um único modelo, deixando uma parte dos resultados inexplicada pela especificação funcional que normalmente levaria a diferentes resultados pontuais (Athey & Imbens, 2015).

Uma solução para tais problemas de estimativas seriam as ferramentas de Machine Learning (ML) como, por exemplo, as técnicas de árvore de decisão, a máquina de suporte vetorial (*support vector machine* – SVM), as redes neurais artificiais (RNAs) e a aprendizagem profunda, dentre outras, que apresentam melhores resultados para modelos mais complexos, concentrando em elevado desempenho computacional, além de lidar com a presença de restrições quanto às relações funcionais lineares ou não (Varian, 2014).

Com essa gama de possibilidades, desenvolveram-se pesquisas utilizando as técnicas de ML para seleção de portfólios (Montenegro & Albuquerque, 2017), analisar a previsão da taxa de câmbio com SVM (Yaohao & Albuquerque, 2019), predição de desempenho de criptomoedas (Yaohao, Albuquerque, Camboim de Sá, Padula & Montenegro, 2018), modelos de precificação de ações e opções (DeSpiegeleer, Madan, Reyners & Schoutens, 2018), construção de modelos de previsão não lineares não paramétricos para risco de crédito (Khandani, Kim & Lo, 2010) e para seleção de gestores financeiros, visto que essa ferramenta serve de apoio à decisão para melhor seleção dos futuros administradores dos fundos (Ludwig & Piovoso, 2005).

As técnicas de aprendizado supervisionado (ML) focam, então, no direcionamento dos modelos a partir de um conjunto de dados (Athey, 2015). Extrapolam, ainda, apresentando resultados mais confiáveis quando os dados são heterogêneos e a forma funcional não pode ser observada. Nesse sentido, os diversos métodos de ML tornam-se mais eficazes para problemas relacionados à predição (Athey & Imbens, 2016), nesse caso, da falência das empresas.

A possibilidade de relações não lineares entre as variáveis usadas constantemente na previsão de falências pode apresentar maior acurácia com as técnicas de ML (Tsai, Hsu & Yen, 2014). Essas variáveis são tratadas como homogêneas e, às vezes, não são, causando riscos de interpretação, principalmente dos efeitos causais e imprecisos. Índices de endividamento, por exemplo, apresentam características distintas quando analisados seus componentes individualmente, explicitando a heterogeneidade entre as empresas, trazendo uma nova perspectiva para estudos que utilizam tais variáveis (Boot & Thakor, 1997; DeMarzo & Fishman, 2007; Park, 2000). A partir disso, suspeita-se que essas características possam ser estendidas para os demais indicadores usados na análise da falência.

A utilização de abordagens não paramétricas, como a floresta causal (FC), facilitaria a compreensão da heterogeneidade, permitindo uma modelagem flexível com elevados níveis de interações e dimensões (Athey & Imbens, 2016; Wager & Athey, 2018). Essa abordagem permite, então, a construção de intervalos de confiança válidos para analisar o tratamento, mesmo considerando um elevado número de variáveis em relação ao tamanho da amostra.

Ganha maior destaque a FC, já que técnicas como K-*nearest neighbor* (KNN) apresentariam limitações quanto ao número de variáveis, elevando o número de dimensões (Zhang & Zhou, 2007), ou seja, maior quantidade de variáveis causaria imprecisão quanto à métrica de distância usada, gerando estimativas imprecisas. Outra opção seria a *long short-term memory* (LSTM), contudo, essa metodologia seria mais indicada em casos de séries temporais longas, já que tem como pressuposto o princípio da evolução temporal das variáveis para a classificação (Hochreiter & Schimidhuber, 1997), não promovendo resultados relevantes nesta pesquisa, já que a série mais longa seria de cinco anos.

Em termos gerais, maximizar a previsibilidade de falências de empresas, principalmente depois de períodos de agravamento, como em uma crise financeira, ganha

maior relevância. Em tais períodos, uma intervenção governamental, por exemplo, auxiliando empresas mais propensas a falir, evitando a diminuição de emprego e renda para a região, seria mais benéfico, diminuindo os efeitos regionais da recessão.

A FC proposta por Athey e Imbens (2016) e Wager e Athey (2018) resolveria, então, esse problema, facilitando as análises. Nessa metodologia, a árvore busca grupos em que os efeitos médios do tratamento mais se diferem. A busca seria por um tratamento individualizado, equilibrando as duas condições. No primeiro momento, a árvore busca encontrar onde os efeitos do tratamento se diferem mais e depois estima os efeitos do tratamento com maior precisão. Além disso, por métodos computacionais, é inserida a condição de honestidade na qual existe uma subdivisão da amostra para treinar a árvore (amostra de treinamento), seguida da aplicação (amostra de validação). Por fim, estima-se em cada uma das folhas, analisando a diferença entre as médias do tratamento e controle, ou seja, a média de se observar uma empresa com características de falida.

É nesse contexto que esta pesquisa busca explorar a metodologia de FC, visando identificar um conjunto de variáveis relevantes referentes à falência de empresas e encontrar padrões de comportamento nos dados de empresas que apresentaram falência. No mais, os modelos mais comuns, a análise discriminante e o *logit*, são os mais usados e, em se tratando de falências, a FC ainda é incipiente, com poucas aplicações, auxiliando os próximos estudos referentes à falência de empresas.

2. REFERENCIAL TEÓRICO

Os estudos sobre falência geram inúmeros resultados relevantes, principalmente sobre a estrutura de capital, indicadores utilizados e sensibilidade do mercado. No que tange à estrutura de capital, a concentração na dívida possibilita menos custos de transação envolvendo a renegociação dos valores. Ao apresentar o plano de recuperação para um volume menor de credores, esses estão mais propensos a aceitar, além de correrem riscos de maiores prejuízos caso ocorra a liquidação. Existe, ainda, a possibilidade de uma alteração da propriedade, resultado na diminuição da credibilidade, aumentando a probabilidade de liquidação (Ivashina, Iverson & Smith, 2016). Ainda no contexto da alavancagem, estruturas mais arriscadas são mais propensas a recorrer a um processo de falência. Essa probabilidade é reduzida quando existe um volume considerado de dívidas com garantias reais (Jostarndt & Sautner, 2010).

Apresentar garantias reais sólidas aos credores, como ativos imobilizados, pode auxiliar na diminuição do processo de falência, já que essas garantias seriam suficientes para honrar as dívidas. Contudo, manter um elevado volume desse tipo de ativo comprometeria a liquidez da empresa. Existe, então, uma relação negativa entre o risco de liquidez e a falência da firma, sendo que tal relação não parece linear (Brogaard, Li & Xia, 2017). Para o cenário italiano, em que o processo de reorganização e liquidação espelha-se nos capítulos 7 e 11 do Título 11 da regulamentação sobre falência e falidos no United States Code (https://uscode.house.gov/browse/prelim@title11&edition=prelim), a empresa, ao enquadrar-se no processo de reorganização, produz aumento nos juros sobre os financiamentos bancários, refletindo diretamente em seus investimentos (Rodano, Serrano-Velarde & Tarantino, 2016).

No que tange aos indicadores de rentabilidade, como o *return on equity* (ROE) e o *return on assets* (ROA), a elevação deste úlimo acima de 15% pode indicar maior propensão a falhas, sendo impulsionada pelo risco do fluxo de caixa combinado com financiamento interno e oneroso. Outros resultados mostraram que a baixa alavancagem representa maior probabilidade de falência, possivelmente refletida pelo baixo volume de crédito (Giordani, Jacobson, Schedvin & Villani, 2014).

O mercado se torna sensível à falência das empresas. O anúncio da falência informa ao mercado a estrutura contábil da firma com dificuldades, bem como seus fluxos de caixa, gerando dois possíveis efeitos: o contágio e o competitivo (Benmelech & Bergman, 2011; Helwege & Zhang, 2016; Hertzel, Li, Officer & Rodgers, 2008; Hertzel & Officer, 2012; Jorion & Zhang, 2007; Lang & Stulz, 1992).

O mercado entende, então, que empresas similares podem estar passando pelos mesmos problemas, sendo esse efeito conhecido como contágio. Em contrapartida, o anúncio da falência transmite informações do quão boas são as empresas restantes, gerando um passo seguinte à expectativa de redistribuição de riqueza no segmento, sendo esse efeito conhecido como competitivo (Lang & Stulz, 1992). Existe, ainda, a possibilidade de efeitos colaterais, reduzindo o valor de ativos similares no mercado secundário, gerando um desequilíbrio na oferta e na demanda (Benmelech & Bergman, 2011).

Há, ainda, a expectativa sobre a sensibilidade do mercado, mostrando que o preço médio das ações das empresas do mesmo segmento apresenta reação negativa, ou seja, queda, podendo ser reflexo do efeito contágio (Lang & Stulz, 1992).

2.1 Falência e ML

Dada a importância da temática falência, os estudos visando à sua previsão cresceram, principalmente nos últimos anos. A comparação de metodologias de ML [SVM, RNA, mínimos quadrados ponderados (MQP) e árvore de decisão, dentre outras] com as metodologias tradicionais (análise discriminante e *logit*) se torna inevitável, com os resultados indicando superioridade para as técnicas computacionais.

Min e Lee (2005) utilizaram o SVM na previsão de falência e foi identificada uma promissora resposta ao comparar as metodologias mais difundidas na literatura, como análise discriminante e *logit*, sendo o SVM superior na capacidade de previsão, uma vez que os parâmetros foram estimados.

No que tange à seleção de indicadores financeiros para a previsão de falência, Yang, You e Ji (2011) utilizaram os MQP e constataram sua superioridade na predição comparada às demais técnicas tradicionais, além de observar a relação complexa e de não linearidade nos parâmetros.

Tsai et al. (2014) compararam diversas metodologias de ML, como árvore de decisão, RNA e SVM, e encontraram que os modelos de ML são superiores em previsibilidade às métricas tradicionais. Dentre essas, o SVM apresentou os melhores resultados se comparado com os demais modelos estudados, apresentando desempenho intermediário. Ao comparar o modelo gaussiano com SVM e o modelo *logit*, encontraram-se melhores previsões com o processo gaussiano do que com SVM e *logit*, e uma leve precisão superior de SVM sobre o *logit* (Antunes, Ribeiro & Pereira, 2017).

Barboza, Kimura e Altman (2017) compararam diversas metodologias com ML e concluíram que essas apresentam melhora substancial na previsão de falência, algo em torno de 10% a mais de precisão, principalmente quando incluem, além das variáveis propostas pelo z *score* de Altman, alguns indicadores financeiros complementares.

Geralmente, se comparadas as metodologias tradicionais com as de ML, existe uma superioridade desta última. Contudo, analisando os resultados dentre as técnicas de ML, as conclusões ainda são contraditórias, dependendo das variáveis utilizadas.

3. METODOLOGIA

Diversos modelos têm sido empregados em finanças visando identificar as próximas empresas a falir. No contexto da análise convencional, os modelos utilizados, análise discriminante (Altman, 1968) e *logit* (Ohlson, 1980), dentre outros, dependem, principalmente, de uma forma funcional preestabelecida pelo pesquisador limitada ao escopo da metodologia. No aprendizado de máquinas, porém, pode haver a extrapolação imposta pelos modelos, chagando a resultados mais satisfatórios.

Para tal, são necessários as variáveis insumo ou independentes – $x \in R$ [rentabilidade, liquidez, alavancagem e produto interno bruto (PIB), dentre outras] – e o resultado ou variável dependente $y \in R$ ou $y \in [0; 1]$ – não falidas ou falidas, tendo como objetivo aprender como os insumos explicam a falência das empresas. Os resultados podem ser modelos não lineares [relação sugerida nos estudos de Giordani et al. (2014) e Brogaard et al. (2017)].

Outras metodologias foram testadas ao longo dos anos, porém, em muitos casos, o foco foi somente a utilização das metodologias, e não uma análise robusta dos resultados encontrados. Uma síntese desses modelos pode ser observada na Tabela 1.

Tabela 1
Alguns modelos usados na predição de falência

Modelo genérico	Modelo específico	Alguns autores que utilizaram
Análise discriminante	Básico	FitzPatrick (1932)
	Multivariado	Altman (1968), Lennox (1999), Min e Lee (2005), Cho, Kim e Bae (2009), Lee e Choi (2013), Barboza et al. (2017), García, Marqués, Sánchez e Ochoa-Domínguez (2017)
Logit	Básico	Ohlson (1980), Lennox (1999), Min e Lee (2005), Cho et al. (2009), Premachandra, Bhabra e Sueyoshi (2009), Tseng e Hu (2010), Antunes et al. (2017), Barboza et al. (2017), García et al. (2017)
	Logit de intervalo quadrático	Tseng e Hu (2010)
Probit	Básico	Zmijewski (1984), Lennox (1999)

Tabela 1
Cont.

Modelo genérico	Modelo específico	Alguns autores que utilizaram
Redes neurais	Básico	Pendharkar (2005), Chauhan, Ravi e Chandra (2009), Cho et al. (2009), Tseng e Hu (2010), Tsai et al. (2014), Barboza et al. (2017)
	Propagação reversa	Lee e Choi (2013)
	Multicamada	Zmijewski (1984), Lennox (1999)
	Rede de função de base radial	Tseng e Hu (2010)
	Wavelet treinada em evolução	Chauhan et al. (2009)
	Modelo interativo com peso*	Cho et al. (2009)
	Variação limiar	Pendharkar (2005)
Árvore de decisão	Básico	Min e Lee (2005), Cho et al. (2009), Tsai et al. (2014)
Máquina de suporte vetorial	Básico	Min e Lee (2005), Yang et al. (2011), Tsai et al. (2014), Antunes et al. (2017), García et al. (2017)
	Linear	Barboza et al. (2017)
	Radial	Barboza et al. (2017)
Análise envoltória de dados	Básico	Cielen, Peeters e Vanhoof (2004), Premachandra et al. (2009), Premachandra, Chen e Watson (2011)
Processo gaussiano	Básico	Antunes et al. (2017)

Nota: *Mantiveram-se as nomenclaturas usadas pelos autores.*
**Cho et al. (2009) criaram o modelo interativo com pesos com base na aplicação de diversas metodologias de predição de falência.*
Fonte: *Elaborada pelos autores.*

No entanto, depara-se diretamente com problemas (i) do elevado volume de dimensões e (ii) da heterogeneidade. Abordagens não paramétricas que buscam analisar os efeitos heterogêneos têm bons desempenhos em aplicações com pequenas quantidades de variáveis (Wager & Athey, 2018). Na literatura de ML, existe uma variedade de métodos eficazes, sendo que os mais populares – árvore de regressão, floresta aleatória e SVM, dentre outros – implicam em modelar relações entre os atributos e os resultados (Athey & Imbens, 2016).

Dentre as possibilidades para analisar o efeito da crise financeira de 2007, uma solução seria a inclusão de uma *dummy* de interação, contudo, os modelos ficaram ainda mais complexos, resultando, nesta pesquisa, em mais de 80 variáveis. A seleção de tais variáveis seria possível por meio do *least absolute shrinkage and selection operator* (Lasso) e do pós-Lasso, como será visto na sequência. Entretanto, depararíamo-nos com modelos lineares, pois seriam estimados por mínimos quadrados ordinários (MQO). Usar o SVM também seria uma opção, mas ficaria limitado à não exploração das características não observadas (particularidades) das empresas. Já a proposta da árvore e da FC se apresentam mais indicadas nesse contexto, pois possibilitariam condições de observar as características de falência mais latente, considerando as particularidades de cada conjunto de empresas.

3.1 Tratamento Condicional

Na literatura de aprendizado de máquinas baseada em predição, a árvore de regressão apresenta características pouco diferentes dos outros métodos, produzindo partições da população baseadas nas variáveis de maneira que todas as unidades de uma partição recebam a mesma previsão (Athey & Imbens, 2016).

A proposta desta pesquisa, então, seria aplicar uma metodologia incipiente no contexto de finanças, principalmente no tocante à avaliação de falências, analisando suas características. Assim, os trabalhos de Athey e Imbens (2016) e Wager e Athey (2018) foram aplicados às FCs.

As FCs são dotadas de propriedades que promovem a imparcialidade e a normalidade assintótica, produzindo a partição da população de acordo com as variáveis em que todas as partições recebem a mesma previsão. Formalizando o problema com base em Athey e Imbens (2016), temos então N unidades com $i = 1..., N$ e existindo um par para cada unidade $Y_i(0); Y_i(1)$, tendo efeito causal dado por $t_i = Y_i(1)\ Y_i(0)$. Denotamos, ainda, um indicador binário $W_i \in \{0,1\}$ com $W_i = 0$, indicando que não recebeu o tratamento, e $W_i = 1$ que recebeu; tem-se, então:

$$Y_i^{obs} = Y_i(W_i) = \begin{cases} Y_i(0)\ se\ W_i = 0 \\ Y_i(1)\ se\ W_i = 1 \end{cases} \quad (1)$$

Tem-se, ainda, X_i como vetor composto por K variáveis não afetadas por esse tratamento, gerando, então, um conjunto de observações compostas por Y_i^{obs}, W_i, X_i com $i = 1,..., N$, sendo uma amostra independente e identicamente distribuída. Assume-se, ainda, que as observações possam ser trocadas e, em um experimento randomizado com probabilidades de atribuições de tratamento constantes, $e(x) = p$ para os valores de x, temos a probabilidade do efeito marginal do tratamento dado por $p = pr(W_i = 1)$ e a do tratamento condicional dado por $e(x) = pr(W_i = 1|X_i = x)$. Logo se chega a:

$$W_i \perp (Y_i(0), Y_i(1))| X_i \quad (2)$$

Assim, tem-se como efeito de tratamento médio condicional (*conditional averange treatment effects* – CATE):

$$\tau(x) \equiv E[Y_i(1) - Y_i(0)|X_i = x] \quad (3)$$

Com isso, Athey e Imbens (2016) conseguiram estimativas mais precisas para o efeito médio condicional do tratamento, ou seja, $\hat{\tau}(.)$, em que $\tau(x)$ é baseado no particionamento dos recursos, não variando nas partições. O tratamento é atribuído aleatoriamente nas subpopulações associadas por $X_i = x$, indicando que, uma vez conhecidas todas as características observáveis do indivíduo i, o *status* do tratamento não gera informação extra sobre seus possíveis resultados.

3.2 Pós-Lasso

Uma simples possibilidade de análise do efeito condicional a respeito de algum tratamento e as interações de seus efeitos podem ser realizadas por meio do Lasso (procedimento adotado para selecionar as variáveis relevantes em um modelo de regressão). Temos, então, o seguinte modelo:

$$Y_i = \alpha + \beta_w W_i + \beta_x X_i + \beta_{xw} X_i W_i + \epsilon_i \quad (4)$$

Logo, se CATE é o verdadeiro modelo, pode-se escrever da seguinte maneira:

$$\tau(x) = E|[Y_i|X = x, W = 1] - E[Y_i|X = x, W = 0] = \beta_w + \beta_{xw} x \quad (5)$$

A equação 5 implica em diferentes subpopulações indexadas por $X_i = x$, tendo efeitos diferentes para $\beta_{xw} \neq 0$. Essa abordagem é bem comum quando as dimensões das variáveis são pequenas ($p = dim(X_i)$), usando MQO. Contudo, o problema aumenta à medida que p cresce e tende a $p > n$, inviabilizando a aplicação de MQO. A solução aceitável seria, então, aplicar o Lasso e posteriormente o pós-Lasso, selecionando as variáveis que melhor explicam a variável dependente por meio de MQO. Esses procedimentos apresentam propriedades vantajosas quando os parâmetros de regularização são escolhidos adequadamente (Belloni, Chernozhukov & Hansen, 2014b; Belloni et al., 2014a), além de apresentarem imparcialidade e normalidade assintótica.

3.3 FC

Com a possibilidade de elevada dimensão, uma solução seria, então, a FC. Em um contexto amplo, as árvores de regressão e as florestas podem ser consideradas vizinhas, usando uma métrica adaptativa nas aproximações. Geralmente, esses tipos de método usam a distância euclidiana para analisar os vizinhos mais próximos. As árvores de decisão podem apresentar folhas mais estreitas ao longo das direções em que o sinal muda rapidamente, e mais largo em outras direções. Assim, pode-se construir uma árvore causal que se assemelhe à árvore de regressão, encontrando um ponto em que a elevada dimensionalidade não cause tanto problema para as estimativas (Wager & Athey, 2018).

Para tal construção, suponha que existam amostras independentes (X_i, Y_i) de uma árvore de regressão. Divide-se, então, o espaço até particioná-lo em um conjunto de folhas L contendo apenas amostras de treinamento. Dado um ponto x, avalia-se o valor de predição $\hat{\mu}(x)$, identificando a folha $L(x)$, a qual contém x, estabelecendo:

$$\hat{\mu}(x) = \frac{1}{|\{i:X_i \in L(x)\}|} \sum_{\{i:X_i \in L(x)\}} Y_i \quad (6)$$

As FCs são adaptativas e flexíveis, tornando-se eficientes para a estimativa de parâmetros locais, como, por exemplo, a aplicação do CATE (Athey, Tibshirani & Wager, 2019). São calculados estimadores ponderados localmente, ou seja, estimam-se os efeitos do tratamento em um alvo específico $X_i = x$, dando maiores pesos para as observações mais relevantes. O principal benefício seria a maior eficiência na escolha das dimensões mais importantes, reduzindo o problema da dimensionalidade. Ao incorporar o tratamento condicional (CATE), temos:

$$\hat{\mu}(x) = \frac{1}{|\{i:W_i=1 \in L(x)\}|} \sum_{\{i:W_i=1 \in L(x)\}} Y_i - \frac{1}{|\{i:W_i=0 \in L(x)\}|} \sum_{\{i:W_i=0 \in L(x)\}} Y_i \quad (7)$$

Tem-se, então, FC gerando um conjunto *B* de árvores causais, em que cada uma produz uma estimativa $\hat{\tau}(x)$. As florestas, então, agregam suas predições calculando a média $B^{-1}\sum_{b=1}^{B}\hat{\tau}_b(x)$. Pela média da saída de muitas árvores, pode-se também calcular o efeito médio do condicional. Esses procedimentos ignoram as informações sobre o resultado, já que colocam divisões de amostra, chamando-as de honestidade, produzindo folhas grandes com normalidade assintótica em cada uma. Ressalta-se, ainda, que nenhum dado foi desperdiçado, satisfazendo as propriedades de honestidade.

As divisões da amostra, também conhecidas como particionamento amostral, são realizadas, gerando uma amostra de estimação e uma de teste. Após esse procedimento, são estimados os resultados e realizado um processo de validação cruzada no qual é possível prever as estimativas pontuais do efeito do tratamento na amostra de estimativa. Ainda nesse procedimento, a árvore é podada baseada em seu nível de complexidade (*complexity parameter*).

Com isso, assume-se que as árvores causais individuais na floresta são subamostras aleatórias de exemplos de treinamento (Athey & Imbens, 2016). Observam-se, ainda, os vários parâmetros de ajuste, como tamanho mínimo de nós para as árvores e validação cruzada, minimizando as perdas e a redução dos erros padrão. A FC pode ser estimada por meio do pacote causalTree proposto por Athey (2019) para o *software* R®. Vide também o *link* do código no Github (https://github.com/susanathey/causalTree). Demais procedimentos e complementos podem ser observados em seu manual. Sugerimos também a leitura de Vapnik (2000) para mais informações sobre ML.

3.4 Dados e Variáveis Utilizados

Para o mercado, seria interessante identificar as empresas antes de apresentar as características de falência, minimizando suas perdas de investimentos. Tais modelos ou metodologias tornam a avaliação imparcial, isenta de influências subjetivas, permitindo, ao analista, classificar os riscos da empresa quanto a seu futuro e capacidade de gerar resultados.

Para tal verificação, as técnicas de previsão de falência são divididas em: análise qualitativa, com modelos subjetivos; análise univariada, usando taxas baseadas em dados contábeis ou indicadores de mercado; análise multivariada, incluindo modelos de análise discriminante, *logit*, *probit*, não lineares, redes neurais, z-*score* de Altman, o-*score* de Ohlson e modelos com base no valor de mercado, dentre outros (Altman & Hotchkiss, 2007). Modelos como os de Altman (1968) usam a análise discriminante para classificar as empresas como solventes e insolventes.

Limitações de tais estudos são constatadas quando pode apresentar relações não lineares entre as variáveis estudadas, como, por exemplo, a falência e os principais indicadores das empresas (alavancagem, rentabilidade, liquidez) (Giordani et al., 2014). Outras limitações são de cunho das modelagens, como a normalidade dos dados utilizados para a análise discriminante, bem como a linearidade das variáveis. Um problema associado à rede neural refere-se à compreensão e resoluções dos padrões encontrados.

Quanto às causas da falência, não existe um fator isolado preponderante da falência das empresas. Os primeiros estudos usaram somente variáveis endógenas, relacionadas aos indicadores de rentabilidade, liquidez e alavancagem (Altman, 1968; Deakin, 1972; Ohlson, 1980). Seguindo uma mesma linha com variáveis internas, Giordani et al. (2014) adotaram a metodologia *logit* padrão aumentado, na qual buscavam entender as relações não lineares das variáveis que influenciam a falência, e encontraram resultados significantes e robustos.

Em um segundo momento, existem os defensores de que a falência das empresas sofre influência externa, ou seja, variáveis exógenas relacionadas à situação econômica do país e às políticas governamentais, já que os indicadores internos não apresentam informações suficientes sobre as condições econômicas enfrentadas pelas empresas (Johnson, 1970). Giordani et al. (2014) também sugerem a inclusão de variáveis externas aos modelos de falência e advertem, ainda, sobre a necessidade de abordagens não lineares.

No que tange às variáveis exógenas, existem argumentos mostrando que empresas menores são mais propensas a falir devido a diversos fatores, como: (i) empresas maiores aparentam ter mais facilidade em aproveitar os efeitos de escala; (ii) empresas maiores têm mais poder de barganha com fornecedores e instituições financeiras, dentre outros; e (iii) empresas maiores tendem a se beneficiar de maior experiência ou aprendizagem (Strömberg, 2000).

Cabe destacar, ainda, que, em algumas situações, é aconselhável construir modelos específicos para o setor, havendo distinção entre o tamanho das empresas (Mensah, 1984; Taffler, 1984). A síntese de alguns estudos e variáveis pode ser observada na Tabela 2.

Tabela 2
Algumas variáveis usadas nos modelos de falência

Variáveis endógenas	Autores
Capital circulante líquido/AT	Beaver (1966), Altman (1968), Deakin (1972), Altman, Haldeman e Narayanan (1977)
Lucros retidos/AT	Altman (1968), Altman et al. (1977), Ohlson (1980)
EBITDA/AT	Deakin (1972)
EBIT/AT	Altman (1968), Altman et al. (1977), Giordani et al. (2014)
Valor de mercado do PL/VCP	Altman (1968)
Vendas/AT	Altman (1968)
Taxa líquida/AT	Beaver (1966), Deakin (1972)
Passivo total/AT	Beaver (1966), Deakin (1972), Ohlson (1980), DeYoung (2003), Jostarndt e Sautner (2010), Giordani et al. (2014)
Ativo circulante/AT	Deakin (1972)
Capital de giro/AT	Deakin (1972), Ohlson (1980), Cole e Gunther (1995)
Caixa/AT	Deakin (1972)
Fluxo de caixa/AT	Beaver (1966)
Ativo circulante/PC	Beaver (1966), Deakin (1972), Altman et al. (1977), Ohlson (1980)
Ativo circulante líquido/PC	Deakin (1972), Giordani et al. (2014)
Caixa/Passivo circulante	Deakin (1972)
Ativo circulante/Vendas	Deakin (1972)
Ativo circulante líquido/Vendas	Deakin (1972)
Caixa/Vendas	Deakin (1972)
Capital de giro/Vendas	Deakin (1972)
Provisões de fundos/AT	Ohlson (1980)
Variáveis exógenas	
Tamanho	Altman et al. (1977), Ohlson (1980), Cole e Gunther (1995), Strömberg (2000), DeYoung (2003), Jostarndt e Sautner (2010), Giordani et al. (2014)
Produto interno bruto	Giordani et al. (2014)
Idade	Jostarndt e Sautner (2010), Giordani et al. (2014)

AT = ativo total; EBIT = earnings before interest and taxes; EBITDA = earnings before interest, taxes, depreciation and amortization; PL = patrimônio líquido; VCP = valor contábil do passivo; PC = passivo circulante; Ln = logaritmo natural.
Fonte: *Elaborada pelos autores.*

Giordani et al. (2014) enfatizam que os indicadores internos são explorados frequentemente nas análises de insolvência, refletindo a estrutura de capital, lucratividade e liquidez das empresas. No que tange à alavancagem, os autores argumentam que, em condições de falência, o passivo supera os ativos. Quanto ao lucro e à liquidez, esses fornecem relevantes informações sobre a escassez de ativos líquidos para dar continuidade às atividades da empresa, com gastos contínuos e pagamento de dívidas.

O capital circulante líquido baixo é problema frequente apresentado por empresas em situação de falência, já que os recursos são consumidos constantemente pelas perdas operacionais, diminuindo a proporção dos ativos correntes, representado, geralmente, pela liquidez da empresa. No que tange aos lucros retidos/ativos totais, indicam que empresas mais novas tendem a ter menores resultados do que empresas consolidadas no mercado. Segundo Altman (1968), essa variável testada individualmente foi a mais relevante para discriminar os grupos entre empresas falidas e não falidas.

A estrutura de endividamento também é relevante para explicar a falência das empresas. Empresas mais endividadas com bancos estão mais propensas a se reestruturarem devido à maior facilidade de renegociação da dívida (Jostarndt & Sautner, 2010). O risco de insolvência das grandes empresas é reduzido devido ao seu grande volume em ativos, ou seja, são grandes demais para falir (Acharya & Mora, 2015), dando maior relevância para a variável Tamanho. A inclusão de variáveis setoriais compensaria as variações econômicas causadas pelas oscilações do mercado, principalmente devido a alguma crise financeira, setorial, tecnológica ou de suprimentos, dentre outras.

4. ANÁLISE DOS DADOS

Nos últimos anos, a literatura de ML trabalhou fortemente na produção de estimativas com qualidade, até mesmo em dados com elevada dimensão. As previsões podem ser utilizadas para direcionar pequenas populações com características específicas, como, por exemplo, falência empresarial. Visando analisar a heterogeneidade entre as empresas no mercado, foram listadas diversas variáveis contábeis e setoriais de 1.247 empresas.

Foram selecionadas 1.247 empresas estadunidenses de 10 setores classificados de acordo com a Thomson Reuters Business Classification. Os balanços selecionados envolvem os cinco anos do processo de falência, já que existe a comprovação de declines nos indicadores (Kalay, Singhal & Tashjian, 2007). Dentre essas empresas, 66 pediram falência, sendo que 22 faliram antes de 2008 e 44 depois – período de tratamento. Para uma mensuração mais próxima, foram coletados os balanços das empresas não falidas no mesmo ano das empresas falidas, totalizando 32.188 observações trimestrais retiradas da base da Thomson Reuters.

Percebe-se o grande desequilíbrio amostral, com 1.181 empresas não falidas e 66 falidas, caracterizando a proporção desigual entre as duas classes (falidas e não falidas). Para resolver esse problema, utilizou-se a metodologia *sythetic minority oversampling technique* (SMOTE). O SMOTE é um algoritmo de geração de dados artificiais para balancear a classe minoritária com base em vizinhos mais próximos. A classe majoritária também é reamostrada, aumentando o volume de dados (Chawla, Bowyer, Hall & Kegelmeyer, 2002).

No que tange às variáveis, na aplicação da metodologia de árvore e FC, bem como na das demais técnicas de ML, seria interessante um maior número de variáveis, tendo em vista captar detalhadamente as características das empresas. Esse processo gera grande dificuldade, já que existem dados ausentes em boa parte dos balanços, comprometendo, assim, um número elevado de observações. Listamos, então, um conjunto de variáveis patrimoniais e setoriais para realizarmos a aplicação da metodologia. A estatística descritiva sem os dados sintéticos pode ser observada na Tabela 3.

Tabela 3
Estatística descritiva dos dados

Nome	Sigla	Média	DP	Mínimo	Mediana	3-quantis	Máximo
Patrimônio líquido	TE_I	0,38	3,06	-272,70	0,61	0,78	9,28
Passivo total	TL_I	0,67	6,36	-0,09	0,39	0,60	730,98
Passivo longo prazo	TLTD_I	0,13	0,42	0,00	0,02	0,16	25,97
Total recebíveis líquidos	TRN_I	0,17	0,15	-0,21	0,14	0,23	10,81
Rendimentos totais	TR_I	0,33	0,86	-6,58	0,27	0,41	134,18
Equipamentos	PPETN_I	0,22	0,23	-0,17	0,15	0,31	11,79
Lucros retidos	RE_AD_I	-3,53	95,27	-16.217,69	-0,06	0,32	4,69
Ativo total	LN_TA	17,76	1,78	5,01	17,90	18,98	33,96
Ativo circulante	TCA_I	0,63	5,86	0,00	0,62	0,78	756,76
Passivo circulante	TCL_I	0,43	2,93	0,00	0,22	0,36	273,70
Débito total	TD_I	0,24	1,14	0,00	0,08	0,27	77,76
Lucro bruto	GP_I	0,11	0,26	-8,50	0,09	0,14	32,79
Lucro após impostos	NIAT_I	-0,07	1,03	-76,59	0,00	0,02	15,27
Vendas líquidas	NS_I	0,33	0,86	-6,58	0,26	0,41	134,18
Dívidas de curto prazo	NPSTD_I	0,05	0,65	0,00	0,00	0,00	59,14
Lucro operacional	OI_I	-0,05	0,86	-76,82	0,01	0,03	15,82
Custos dos produtos	CR_I	0,21	0,65	-5,69	0,15	0,27	101,39
EBIT	EBIT_I	-0,06	1,74	-175,93	0,01	0,03	15,82
EBITDA	EBITDA_I	-0,04	1,67	-165,73	0,02	0,04	15,82
Contas a pagar	AP_I	0,13	0,81	0,00	0,06	0,12	92,98
Despesas provisionadas	AE_I	0,11	0,65	-21,24	0,06	0,10	73,80
Caixas e equivalentes	CSTI_I	0,23	0,23	-0,01	0,15	0,37	3,61
Estoque	CST_I	0,23	0,23	-0,01	0,15	0,37	3,61

Tabela 3
Cont.

Nome	Sigla	Média	DP	Mínimo	Mediana	3-quantis	Máximo
***Dummies* setoriais**							
Tecnologia	D_T	0,24	0,43	0,00	0,00	0,00	1,00
Materiais básicos	D_BM	0,05	0,23	0,00	0,00	0,00	1,00
Consumo cíclico	D_CC	0,18	0,39	0,00	0,00	0,00	1,00
Consumo não cíclico	D_CNC	0,05	0,22	0,00	0,00	0,00	1,00
Energia	D_E	0,06	0,25	0,00	0,00	0,00	1,00
Financeiro	D_F	0,02	0,14	0,00	0,00	0,00	1,00
Hospitalar	D_H	0,17	0,37	0,00	0,00	0,00	1,00
Indústria	D_I	0,19	0,39	0,00	0,00	0,00	1,00
Telecomunicações	D_TS	0,01	0,12	0,00	0,00	0,00	1,00
Utilidades	D_U	0,01	0,08	0,00	0,00	0,00	1,00

Nota: *Valores em percentuais. Todas as variáveis foram ponderadas pelo ativo total. Para a variável ativo total, usou-se o logaritmo natural.*
DP = desvio padrão; EBIT = earnings before interest and taxes; EBITDA = earnings before interest, taxes, depreciation and amortization.
Fonte: *Elaborada pelos autores.*

Como esperado, existe uma grande variação entre as empresas, principalmente no tamanho. Essa variação contribui bastante para a heterogeneidade das empresas. Observa-se, ainda, que, apesar de haver muitas variáveis contábeis, há baixa correlação entre elas (Figura 1).

A variável *X29* refere-se à variável binária, indicando falidas ou não falidas, e a variável *TRA* refere-se à variável binária de tratamento – antes e depois da crise. Cabe destacar que não estamos interessados nos efeitos causais alcançados a partir, principalmente, de métricas paramétricas, mas em analisar algumas variáveis que possam indicar partições relevantes para indicar a solidez de uma empresa. Nesse contexto, os resultados da FC não podem ser interpretados como efeitos parciais, mantendo as demais variáveis constantes.

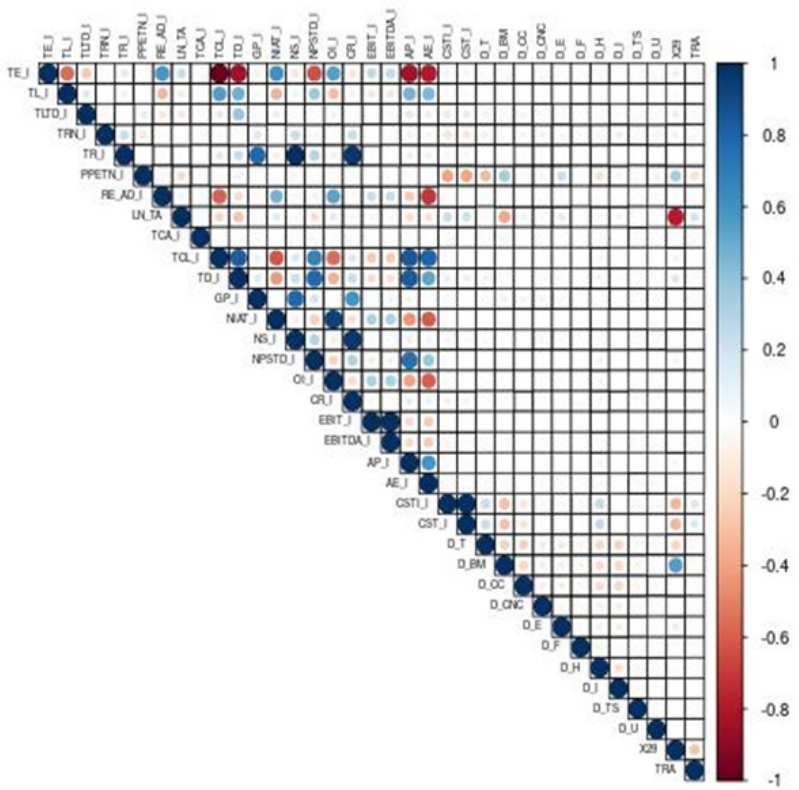

Figura 1 *Correlação entre as variáveis*
Fonte: *Elaborada pelos autores.*

4.1 Análise Pós-Lasso

Uma maneira simples de analisar os efeitos causais entre as variáveis antes e pós-colapso financeiro seria por meio de interações simples com um modelo linear, como descrito na equação 4. Athey e Imbens (2016) advertem que tal metodologia seria relevante em modelos com poucas variáveis, tornando-se um problema quando existe elevado volume. Com elevada dimensão, uma solução seria realizar o Lasso como uma espécie de operador para seleção de variáveis relevantes ao modelo (Athey, Imbens, Pham & Wager, 2017) e depois aplicar a regressão por MQO (Belloni et al., 2014b). Realizados tais procedimentos, os resultados podem ser observados na Tabela 4.

Tabela 4
Resultado pós-Lasso

Variáveis	Estimativas	Erro padrão	Pr(>|t|)	Variáveis	Estimativas	Erro padrão	Pr(>|t|)
(Intercepto)	1,11451	0,00129	0,00000	I(TE_I * W)	0,11359	0,00176	0,00000
TL_I	0,00088	0,00010	0,00000	I(TRN_I * W)	-0,01963	0,00547	0,00033
TLTD_I	0,15660	0,00199	0,00000	I(PPETN_I * W)	0,06758	0,00502	0,00000
PPETN_I	0,21366	0,00350	0,00000	I(LN_TA * W)	-0,01647	0,00013	0,00000
RE_AD_I	0,00021	0,00001	0,00000	I(TCA_I * W)	0,00220	0,00023	0,00000
TCA_I	-0,00169	0,00015	0,00000	I(TCL_I * W)	0,12330	0,00187	0,00000
OI_I	0,00277	0,00163	0,09050	I(GP_I * W)	-0,05196	0,00370	0,00000
CR_I	-0,00059	0,00139	0,67372	I(NPSTD_I * W)	0,02562	0,00109	0,00000
EBITDA_I	0,00006	0,00048	0,90396	I(OI_I * W)	-0,01134	0,00189	0,00000
AP_I	-0,01409	0,00108	0,00000	I(AE_I * W)	-0,00835	0,00219	0,00013
D_BM	0,56272	0,00166	0,00000	I(CSTI_I * W)	0,12206	0,00382	0,00000
D_CC	0,11772	0,00148	0,00000				
D_CNC	0,13174	0,00248	0,00000				
D_TS	0,13751	0,00470	0,00000				

Lasso = least absolute shrinkage and selection operator.
Fonte: *Elaborada pelos autores.*

Com as interações, o modelo ficaria com 66 variáveis, das quais 33 são as iniciais do modelo (33 variáveis, sendo 23 contábeis e 10 setoriais) e 33 interações. Observa-se que o volume de interações relevantes *I*(**W*), principalmente nas variáveis internas da empresa, é elevado, 11 no total. Os indicativos setoriais *D* só foram relevantes em quatro ocasiões, evidenciando que, antes da crise financeira, os setores de Materiais básicos (*D_BM*), Consumo cíclico (*D_CC*), Consumo não cíclico (*D_CNC*) e Telecomunicações (*D_TS*) foram mais afetados nos processos de falência. Após a crise, os resultados seriam amplos, não havendo interações relevantes. Porém, existe uma limitação quanto à interpretação desse modelo, já que se trata de uma regressão linear.

Esses resultados são bem genéricos quanto a uma possível previsibilidade, já que se encontram efeitos diferentes nas mais diversas empresas. Dadas as características individuais de cada empresa, a possibilidade de renegociação de dívidas, por exemplo, causaria distorções quanto às possibilidades de intervenção nas empresas. Outro ponto relevante seriam as características do ativo circulante quanto à liquidez seca e imediata e giro. Os resultados operacionais e não operacionais, bem como a qualidade do lucro envolvido, podem ser determinantes relevantes para uma empresa falir ou não. E com esses resultados (Tabela 4), as variáveis são tratadas homogeneamente.

4.2 Análise do Tratamento Condicional e Árvore Causal

Nesse cenário, existe a necessidade de saber em quais subpopulações a crise financeira teve maior efeito. Athey e Imbens (2016) afirmam que, nesses casos, uma maneira orientada por dados para identificar a heterogeneidade relevante pode ser conveniente. As árvores causais, então, produzem esse indicativo a partir dos dados para entender a heterogeneidade e onde ela está de acordo com o espaço de cada variável, gerando estimativas imparciais do tratamento em cada subgrupo. A árvore inicial foi gerada com 294 folhas. O erro de validação cruzada (*x-val*) nem sempre reduz quando a árvore se torna complexa (para ficar de fácil compreensão, usa-se uma analogia ao modelo de regressão: com a inclusão de mais variáveis ao modelo, seu poder de predição não aumenta). Um bom ponto de corte seria quando os pontos cortam e

se posicionam abaixo da linha horizontal, optando pelo ponto mais à esquerda, geralmente sendo o menor valor de *xerror*. Após todos esses procedimentos de análise, o parâmetro de regularização converge em 156 divisões – o valor de *xerror* deixa de diminuir.

Sabe-se, ainda, que os coeficientes de interação gerados são os efeitos médios de tratamento em cada uma das folhas (Tabela 5). A árvore, então, depois dos ajustes, passaria a ter 156 folhas. Sabe-se, ainda, que em todas essas folhas os tratamentos são relevantes.

Tabela 5
Efeito do tratamento por folha

Folha	Estimativa	Folha	Estimativa	Folha	Estimativa	Folha	Estimativa
Folha_1	-1	Folha_40	-0,89796	Folha_79	-0,39655	Folha_118	0,38961
Folha_2	-0,99813	Folha_41	-0,89305	Folha_80	-0,34615	Folha_119	0,39011
Folha_3	-0,99448	Folha_42	-0,89286	Folha_81	-0,34211	Folha_120	0,40705
Folha_4	-0,99375	Folha_43	-0,88889	Folha_82	-0,3125	Folha_121	0,45554
Folha_5	-0,99058	Folha_44	-0,88462	Folha_83	-0,30303	Folha_122	0,53782
Folha_6	-0,98944	Folha_45	-0,88413	Folha_84	-0,30189	Folha_123	0,54167
Folha_7	-0,98936	Folha_46	-0,875	Folha_85	-0,29412	Folha_124	0,56061
Folha_8	-0,98924	Folha_47	-0,87097	Folha_86	-0,27778	Folha_125	0,6
Folha_9	-0,98221	Folha_48	-0,86957	Folha_87	-0,27451	Folha_126	0,62372
Folha_10	-0,98077	Folha_49	-0,86538	Folha_88	-0,2525	Folha_127	0,63333
Folha_11	-0,97857	Folha_50	-0,86207	Folha_89	-0,25	Folha_128	0,66885
Folha_12	-0,97619	Folha_51	-0,86	Folha_90	-0,22222	Folha_129	0,74868
Folha_13	-0,97464	Folha_52	-0,85185	Folha_91	-0,2151	Folha_130	0,78481
Folha_14	-0,97297	Folha_53	-0,84783	Folha_92	-0,14474	Folha_131	0,79081
Folha_15	-0,96985	Folha_54	-0,84328	Folha_93	-0,14444	Folha_132	0,79094
Folha_16	-0,9697	Folha_55	-0,84	Folha_94	-0,05294	Folha_133	0,8
Folha_17	-0,96769	Folha_56	-0,83871	Folha_95	-0,04819	Folha_134	0,8125
Folha_18	-0,96636	Folha_57	-0,81731	Folha_96	-0,04762	Folha_135	0,81818
Folha_19	-0,96592	Folha_58	-0,78571	Folha_97	-0,03509	Folha_136	0,82467
Folha_20	-0,96552	Folha_59	-0,78182	Folha_98	-0,02817	Folha_137	0,82524
Folha_21	-0,96226	Folha_60	-0,75177	Folha_99	-0,02542	Folha_138	0,82927
Folha_22	-0,96104	Folha_61	-0,75	Folha_100	-0,01471	Folha_139	0,83888
Folha_23	-0,95808	Folha_62	-0,67701	Folha_101	-0,01407	Folha_140	0,84615
Folha_24	-0,95288	Folha_63	-0,67059	Folha_102	-0,01316	Folha_141	0,86194
Folha_25	-0,94767	Folha_64	-0,66667	Folha_103	-0,00855	Folha_142	0,86517
Folha_26	-0,94643	Folha_65	-0,66365	Folha_104	-0,00131	Folha_143	0,86842
Folha_27	-0,9449	Folha_66	-0,65625	Folha_105	-0,00031	Folha_144	0,89836
Folha_28	-0,93023	Folha_67	-0,65476	Folha_106	-0,00136	Folha_145	0,90398
Folha_29	-0,9292	Folha_68	-0,65385	Folha_107	0,00201	Folha_146	0,91525
Folha_30	-0,92593	Folha_69	-0,64706	Folha_108	0,00678	Folha_147	0,92011
Folha_31	-0,92126	Folha_70	-0,62805	Folha_109	0,00797	Folha_148	0,92537
Folha_32	-0,92	Folha_71	-0,6156	Folha_110	0,01109	Folha_149	0,94
Folha_33	-0,91824	Folha_72	-0,58696	Folha_111	0,01667	Folha_150	0,94231
Folha_34	-0,91701	Folha_73	-0,58283	Folha_112	0,02041	Folha_151	0,95187
Folha_35	-0,91667	Folha_74	-0,56604	Folha_113	0,0303	Folha_152	0,9575
Folha_36	-0,91463	Folha_75	-0,53061	Folha_114	0,05085	Folha_153	0,96364
Folha_37	-0,91183	Folha_76	-0,43836	Folha_115	0,15094	Folha_154	0,96923
Folha_38	-0,9108	Folha_77	-0,42	Folha_116	0,24316	Folha_155	0,975
Folha_39	-0,90691	Folha_78	-0,40789	Folha_117	0,30556	Folha_156	1

Nota: *O erro padrão tem como valor máximo e mínimo 0,03762 e 0,00075, respectivamente.*
Fonte: *Elaborada pelos autores.*

As análises são similares a uma regressão MQO. Observe que os dados estão em ordem decrescente e somente a partir da folha 107 os coeficientes são positivos; assim, a crise teria efeito negativo em mais da metade das folhas, mostrando sua relevância para as variáveis contábeis analisadas.

Dadas as condições empresariais e suas particularidades, a crise financeira ocorrida afetou diferentemente as diversas empresas, já que o efeito do tratamento é diferente em cada uma das folhas, calculado por meio do teste *F*. Cabe destacar, ainda, que se uma divisão não ocorreu em uma variável específica, não significa sua irrelevância. Existem diversas maneiras de escolher uma subamostra com os mais diversos efeitos de tratamento, podendo ser altos ou baixos.

O efeito médio geral (média das variáveis) pode ser observado na Tabela 6. As variáveis setoriais, como destacado, foram as que apresentaram um tratamento médio próximo de 0 para as diversas folhas da árvore, indicando menor heterogeneidade. Destacam-se como os setores mais afetados o de Materiais básicos (*D_BM*) e o de Consumo cíclico (*D_CC*), com maior relevância em momentos de crises, sendo esses os setores mais preponderantes para a falência das empresas depois do período de crise.

Tabela 6
Média geral por variável

Nome	Sigla	Média	Nome	Sigla	Média
Patrimônio líquido	TE_I	-0,031596154	EBIT	EBIT_I	-0,068320513
Passivo total	TL_I	1,085089744	EBITDA	EBITDA_I	-0,050916667
Passivo longo prazo	TLTD_I	0,233576923	Contas a pagar	AP_I	0,188685897
Total recebíveis líquidos	TRN_I	0,14500641	Despesas provisionadas	AE_I	0,133032051
Rendimentos totais	TR_I	0,388935897	Caixas e equivalentes	CSTI_I	0,120948718
Equipamentos	PPETN_I	0,309858974	Estoque	CST_I	0,120948718
Lucros retidos	RE_AD_I	-3,572378205	Tecnologia	D_T	0,102634615
Ativo total	LN_TA	15,31235897	Materiais básicos	D_BM	0,191153846
Ativo circulante	TCA_I	0,488692308	Consumo cíclico	D_CC	0,312897436
Passivo circulante	TCL_I	0,697269231	Consumo não cíclico	D_CNC	0,097602564
Débito total	TD_I	0,518230769	Energia	D_E	0,029730769
Lucro bruto	GP_I	0,108211538	Financeiro	D_F	0,004371795
Lucro após impostos	NIAT_I	-0,120929487	Hospitalar	D_H	0,091384615
Vendas líquidas	NS_I	0,389647436	Indústria	D_I	0,148961538
Dívidas de curto prazo	NPSTD_I	0,118884615	Telecomunicações	D_TS	0,020955128
Lucro operacional	OI_I	-0,094794872	Utilidades	D_U	0,000269231
Custos dos produtos	CR_I	0,281365385			

EBIT = earnings before interest and taxes; EBITDA = earnings before interest, taxes, depreciation and amortization.
Fonte: *Elaborada pelos autores.*

Empresas que atuam em setores como Utilidades, Financeiros, Telecomunicações, Energia, Hospitalar, Consumos não cíclicos e Tecnologia são as menos afetadas pela crise financeira, possivelmente devido à necessidade dos itens produzidos. No que tange às variáveis utilizadas, observa-se que as mais afetadas seriam o Patrimônio líquido, EBITDA, EBIT, o Lucro operacional, Lucro após os impostos e Lucros retidos. Como esperado, as variáveis de Lucro e Patrimônio líquido tiveram os efeitos negativos com médias de tratamento inferiores a 0, com destaque para lucros retidos com o menor coeficiente.

Devido ao tamanho da árvore estimada, que ficaria invisível nesse documento, não seria possível incorporar a figura, mas o principal ponto de segregação seria o tipo de setor do qual as empresas fazem parte. Destaca-se, como primeira divisão, o setor de Materiais básicos (*D_BM*) e, para determinados volumes em ativos, menores empresas ($LN\_TA < e^{12,238}$), a próxima divisão seria Lucros retidos. Para empresas que não pertencem ao setor de Materiais básicos (< 0,5), a partição seguinte seria nos Ativos totais (*LN_AT*), sendo que, para os maiores que $LN\_TA\ e^{12,238}$, a segregação seria o setor de Consumo cíclico (*D_CC*), ressaltando que empresas de maior porte tentem ser menos afetadas, apresentado elevado volume de subdivisões.

Características como o Total de recebíveis líquidos (*TRN_I*) mostram-se relevantes, tendo em vista a necessidade de aumento dos fluxos de caixa das empresas, principalmente em momentos de recessão. Empresas com *TRN_I*, por exemplo, maior que 16% tenderiam a ter pontos de falência, dependendo do seu tamanho (*LN_AT*) e volume de endividamento (*TL_I*).

Não muito distante do apresentado por Giordani et al. (2014), o tamanho da empresa foi relevante nas principais partições encontradas, amenizado pelo seu elevado volume em ativos, já que empresas menores tendem a ser mais propensas à falência. Existe, ainda, a possibilidade de mais benefícios e intervenções governamentais, visando amenizar o volume em desemprego gerado pela falência das grandes empresas.

Outra variável importante seria a de Vendas líquidas convergindo com um dos indicadores propostos por Altman (1968), mostrando que empresas com mais capacidade de geração de receitas apresentam menos problemas em períodos de crise. As variáveis de liquidez também foram relevantes, bem como os indicativos de rentabilidade.

4.3 FCs

As FCs tornam-se, então, um método adaptativo e eficiente para estimar parâmetros que podem ser definidos por condições locais, como, por exemplo, após a aplicação do CATE. As previsões da FC são estimativas médias de árvore causal, ou seja, estimam-se, no mínimo, duas árvores causais e, após, combinam-se as árvores, gerando as estimativas da FC. Os pesos encontrados em cada uma das folhas das árvores causais evidenciam maior confiabilidade no volume de dimensões importantes, além de serem adaptativos, tornando as estimativas mais robustas diante da heterogeneidade das empresas.

Ao prever as estimativas do CATE e sua variação para cada observação, constata-se pouca variabilidade, com média geral próxima de 0 (Tabela 7) nas linhas Predições e Variância estimada. O termo "Erro viesado", na linha, indica que o erro é devido apenas à variabilidade da amostra dos dados, ou seja, representa o erro que se espera com a construção de uma floresta contendo um número infinito de árvores. Note-se, com isso, a consistência das estimativas com o erro próximo de 0.

Tabela 7
Média geral do efeito de tratamento médio condicional (conditional averange treatment effects – CATE)

	Média	DP	Mínimo	1º quartil	Mediana	3º quartil	Máximo
Média geral do CATE da amostra de teste							
Predições	0	0,04	-1,48	0	0	0	0,68
Variância estimada	0	0,01	0,00	0	0	0	2,98
Erro viesado	0	0,00	0,00	0	0	0	0,27
Média geral do CATE da amostra de validação							
Predições	0	0,04	-1,28	0	0	0	0,68
Variância estimada	0	0,01	0,00	0	0	0	0,64

DP = desvio padrão.
Fonte: *Elaborada pelos autores.*

A partir das previsões do conjunto de teste, estimamos as previsões para a amostra de validação na Tabela 7. Como esperado, as estimativas apresentaram variações bem pequenas, todas próximas de 0, indicando que o modelo se ajusta bem aos parâmetros e aos dados. Portanto, os resultados convergem para a maior possibilidade de previsibilidade, além de tratar homogeneamente as características das empresas analisadas. Constata-se, ainda, a redução no valor máximo da variância estimada, reduzindo do patamar anterior de 2,98 para 0,64. A variável Erro viesado não aparece, pois foi testada na amostra de validação.

As variáveis mais usadas na partição da árvore podem ser vistas na Tabela 8. Contudo, não podemos cair na armadilha de que, com pouca frequência na utilização nas partições, a variável não seja relevante. Observe que a frequência da variável setorial *D_BM* é 0,2%, porém, a principal partição da árvore encontra-se nessa variável.

Tabela 8
Variáveis mais usadas na partição

Variável	Frequência	Variável	Frequência	Variável	Frequência
GP_I	0,26701	OI_I	0,01747	AE_I	0,00589
AP_I	0,14584	TL_I	0,01334	NS_I	0,00516
EBIT_I	0,09351	RE_AD_I	0,01318	D_BM	0,00255
EBITDA_I	0,07489	TCA_I	0,01260	D_I	0,00209
TLTD_I	0,05646	CR_I	0,00979	D_TS	0,00130
PPETN_I	0,04446	CST_I	0,00881	D_CC	0,00038
TE_I	0,04394	NPSTD_I	0,00849	D_T	0,00012
TRN_I	0,04311	CSTI_I	0,00780	D_CNC	0,00012
LN_TA	0,04266	TD_I	0,00709	D_E	0,00000
D_H	0,03818	TR_I	0,00678	D_F	0,00000
NIAT_I	0,02083	TCL_I	0,00614	D_U	0,00000

Fonte: *Elaborada pelos autores.*

Nas subpartições, a variável Lucro bruto (GP_I) foi a que apresentou maior frequência quando se dividiu a árvore, com aproximadamente 27% das aparições. A variável Contas a pagar (AP_I) apresenta-se relevante no processo de determinação de falência das empresas, já que impacta diretamente nos seus fluxos de caixa, bem como na sua credibilidade. Cabe destacar, ainda, que se duas variáveis forem altamente correlacionadas, pode haver particionamento em uma das variáveis, mas não em outra. No entanto, caso haja remoção de uma, a subdivisão pode ocorrer na que sobrou, mantendo as definições em cada folha inalterada.

5. CONSIDERAÇÕES FINAIS

Os resultados indicaram que existem diversas variáveis que normalmente não são incluídas nos modelos de análise e previsão de falências. A variável Vendas líquidas (*NS_I*), conforme apontado por Altman (1968), continua sendo relevante. Destaca-se a importância da inclusão de variáveis indicativas de setores de atuação. Pode-se especular que existem setores mais propensos à falência, principalmente em momentos de crise. Nesta pesquisa, o mais afetado foi o de Materiais básicos (D_BM), que incluem empresas químicas, de exploração mineral e ambientais (papel, madeira e recipientes). Caso não pertença à *D_BM*, outro setor bem afetado seria o de Consumo cíclico (*D_CC*) (automóveis, material de construção, utensílios domésticos, hotéis, produção e entretenimento).

Observamos, ainda, a presença da heterogeneidade entre as empresas que, em muitos casos, são tratadas como idênticas. Os índices de endividamento, por exemplo, em modelos lineares são tratados como similares entre as empresas e não são, dados o tamanho e a capacidade de barganha com fornecedores e governo, entre outros.

Empresas de menor porte também podem apresentar menos capacidade de obtenção de crédito, exigindo, dos gestores, maiores volumes em caixa ou equivalentes para manterem-se funcionando. Com isso, tendem a apresentar maiores indicadores de liquidez. A depender do segmento, as empresas podem apresentar maiores volumes em ativos imobilizados, reduzindo os índices de liquidez; em contrapartida apresentam maiores volumes em depreciação. Essas características devem ser levadas em consideração no tratamento ou intervenção, principalmente em períodos de crise, cabendo aos intervencionistas adotar a melhor estratégia para cada empresa.

Uma limitação dessa metodologia seria a necessidade de uma abordagem quase experimental, necessitando de uma base antes e depois de um fenômeno específico. Analisar sem a necessidade desse evento proporcionaria maior contribuição acadêmica. Sugere-se, como pesquisas futuras, a exploração das características não observadas das empresas por meio de outras metodologias, abordando, por exemplo, o impacto intertemporal nas empresas e nas variáveis, já que essa metodologia proposta não abordaria tais efeitos e suas magnitudes.

REFERÊNCIAS

Acharya, V. V., & Mora, N. (2015). A crisis of banks as liquidity providers. *Journal of Finance*, *70*(1), 1-43. https://doi.org/10.1111/jofi.12182

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

Altman, E. I., Haldeman, R. G., & Narayanan, P. (1977). ZETA analysis: A new model to identify bankruptcy risk of corporations. *Journal of Banking and Finance*, *1*(1), 29–54. https://doi.org/10.1016/0378-4266(77)90017-6

Altman, E. I., & Hotchkiss, E. (2007). *Corporate financial distress and bankruptcy* (3a ed.). Hoboken, NJ: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118267806

Antunes, F., Ribeiro, B., & Pereira, F. (2017). Probabilistic modeling and visualization for bankruptcy prediction. *Applied Soft Computing Journal*, *60*, 831-843. https://doi.org/10.1016/j.asoc.2017.06.043

Athey, S. (2015). Machine learning and causal inference for policy evaluation. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD'15* (p. 5-6). New York, NY. https://doi.org/10.1145/2783258.2785466

Athey, S. (2019). *CausalTree*. Recuperado de https://github.com/susanathey/causalTree

Athey, S., & Imbens, G. (2015). Machine learning methods in economics and econometrics: A measure of robustness to misspecification. *American Economic Review*, *105*(5), 476-480. https://doi.org/10.1257/aer.p20151020

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353-7360. https://doi.org/10.1073/pnas.1510489113

Athey, S., Imbens, G., Pham, T., & Wager, S. (2017). Estimating average treatment effects: Supplementary analyses and remaining challenges. *American Economic Review*, *107*(5), 278-281. https://doi.org/10.1257/aer.p20171042

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148-1178. https://doi.org/10.1214/18-AOS1709

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

Belloni, A., Chernozhukov, V., & Hansen, C. (2014a). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, *28*(2), 29-50. https://doi.org/10.1257/jep.28.2.29

Belloni, A., Chernozhukov, V., & Hansen, C. (2014b). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, *81*(2), 608-650. https://doi.org/10.1093/restud/rdt044

Benmelech, E., & Bergman, N. K. (2011). Bankruptcy and the collateral channel. *Journal of Finance*, *66*(2), 337-378. https://doi.org/10.1111/j.1540-6261.2010.01636.x

Boot, A. W. A., & Thakor, A. V. (1997). Financial system architecture. *Review of Financial Studies*, *10*(3), 693-733. https://doi.org/10.1093/rfs/10.3.693

Brogaard, J., Li, D., & Xia, Y. (2017). Stock liquidity and default risk. *Journal of Financial Economics*, *124*(3), 486-502. https://doi.org/10.1016/j.jfineco.2017.03.003

Chauhan, N., Ravi, V., & Chandra, D. K. (2009). Differential evolution trained wavelet neural networks: Application to bankruptcy prediction in banks. *Expert Systems With Applications*, *36*(4), 7659-7665. https://doi.org/10.1016/j.eswa.2008.09.019

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*(1), 321-357. https://doi.org/10.1613/jair.953

Cho, S., Kim, J., & Bae, J. K. (2009). An integrative model with subject weight based on neural network learning for bankruptcy prediction. *Expert Systems With Applications*, *36*(1), 403-410. https://doi.org/10.1016/j.eswa.2007.09.060

Cielen, A., Peeters, L., & Vanhoof, K. (2004). Bankruptcy prediction using a data envelopment analysis. *European Journal of Operational Research*, *154*(2), 526-532. https://doi.org/10.1016/S0377-2217(03)00186-3

Cole, R. A., & Gunther, J. W. (1995). Separating the likelihood and timing of bank failure. *Journal of Banking and Finance*, *19*(6), 1073–1089. https://doi.org/10.1016/0378-4266(95)98952-M

Deakin, E. B. (1972). A discriminant analysis of predictors of business failure. *Journal of Accountin Research*, *10*(1), 167-179. Retrieved from http://www.jstor.org/stable/2490225

DeMarzo, P. M., & Fishman, M. J. (2007). Optimal long-term financial contracting. *Review of Financial Studies*, *20*(6), 2079-2128. https://doi.org/10.1093/rfs/hhm031

DeSpiegeleer, J., Madan, D. B., Reyners, S., & Schoutens, W. (2018). Machine learning for quantitative finance: Fast derivative pricing, hedging and fitting. *Quantitative Finance*, *18*(10), 1635-1643. https://doi.org/10.1080/14697688.2018.1495335

DeYoung, R. (2003). The failure of new entrants in commercial banking markets: A split-population duration analysis. *Review of Financial Economics*, *12*(1), 7–33. https://doi.org/10.1016/S1058-3300(03)00004-1

FitzPatrick, P. J. (1932). *A comparison of the ratios of successful industrial enterprises with those of failed companies.* Recuperado de https://www.worldcat.org/title/comparison-of-the-ratios-of-successful-industrial-enterprises-with-those-of-failed-companies/oclc/6284198

García, V., Marqués, A. I., Sánchez, J. S., & Ochoa-Domínguez, H. J. (2017). Dissimilarity-based linear models for corporate bankruptcy prediction. *Computational Economics*, *53*, 1019-1031. https://doi.org/10.1007/s10614-017-9783-4

Giordani, P., Jacobson, T., Schedvin, E. Von, & Villani, M. (2014). Taking the Twists into account: Predicting firm bankruptcy

risk with splines of financial ratios. *Journal of Financial and Quantitative Analysis*, *49*(4), 1071-1099. https://doi.org/10.1017/S0022109014000623

Helwege, J., & Zhang, G. (2016). Financial firm bankruptcy and contagion. *Review of Finance*, *20*(4), 1321-1362. https://doi.org/10.1093/rof/rfv045

Hertzel, M. G., Li, Z., Officer, M. S., & Rodgers, K. J. (2008). Inter-firm linkages and the wealth effects of financial distress along the supply chain. *Journal of Financial Economics*, *87*(2), 374-387. https://doi.org/10.1016/j.jfineco.2007.01.005

Hertzel, M. G., & Officer, M. S. (2012). Industry contagion in loan spreads. *Journal of Financial Economics*, *103*(3), 493-506. https://doi.org/10.1016/j.jfineco.2011.10.012

Hochreiter, S., & Schimidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Ivashina, V., Iverson, B., & Smith, D. C. (2016). The ownership and trading of debt claims in Chapter 11 restructurings. *Journal of Financial Economics*, *119*(2), 316-335. https://doi.org/10.1016/j.jfineco.2015.09.002

Johnson, C. G. (1970). Ratio Stability and corporate failure. *The Journal of Finance*, *25*(5), 1166-1168. https://doi.org/10.2307/2325590

Jorion, P., & Zhang, G. (2007). Good and bad credit contagion: Evidence from credit default swaps. *Journal of Financial Economics*, *84*(3), 860-883. https://doi.org/10.1016/j.jfineco.2006.06.001

Jostarndt, P., & Sautner, Z. (2010). Out-of-court restructuring versus formal bankruptcy in a non-interventionist bankruptcy setting. *Review of Finance*, *14*(4), 623-668. https://doi.org/10.1093/rof/rfp022

Kalay, A., Singhal, R., & Tashjian, E. (2007). Is Chapter 11 costly? *Journal of Financial Economics*, *84*(3), 772-796. https://doi.org/10.1016/j.jfineco.2006.04.001

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, *34*(11), 2767-2787. https://doi.org/10.1016/j.jbankfin.2010.06.001

Lang, L. H. P., & Stulz, R. (1992). Contagion and competitive intra-industry effects of bankruptcy announcements. An empirical analysis. *Journal of Financial Economics*, *32*(1), 45-60. https://doi.org/10.1016/0304-405X(92)90024-R

Lee, S., & Choi, W. S. (2013). A multi-industry bankruptcy prediction model using back-propagation neural network and multivariate discriminant analysis. *Expert Systems with Applications*, *40*(8), 2941-2946. https://doi.org/10.1016/j.eswa.2012.12.009

Lennox, C. (1999). Identifying failing companies: A re-evaluation of the logit, probit and DA approaches. *Journal of Economics and Business*, *51*, 347-364.

Ludwig, R. S., & Piovoso, M. J. (2005). A comparison of machine-learning classifiers for selecting money managers. *Intelligent Systems in Accounting, Finance and Management*, *13*(3), 151-164. https://doi.org/10.1002/isaf.262

Mensah, Y. M. (1984). An examination of the stationarity of multivariate bankruptcy prediction models: A methodological study. *Journal of Accounting Research*, *22*(1), 380. https://doi.org/10.2307/2490719

Min, J. H., & Lee, Y. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications*, *28*(4), 603-614. https://doi.org/10.1016/j.eswa.2004.12.008

Montenegro, M. R., & Albuquerque, P. H. M. (2017). Wealth management: Modeling the nonlinear dependence. *Algorithmic Finance*, *6*(1-2), 51-65. https://doi.org/10.3233/AF-170203

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, *18*(1), 109. https://doi.org/10.2307/2490395

Park, C. (2000). Monitoring and structure of debt contracts. *The Journal of Finance*, *55*(5), 2157-2195. https://doi.org/10.1111/0022-1082.00283

Pendharkar, P. C. (2005). A threshold-varying artificial neural network approach for classification and its application to bankruptcy prediction problem. *Computers & Operations Research*, *32*(10), 2561-2582. https://doi.org/10.1016/j.cor.2004.06.023

Premachandra, I. M., Bhabra, G. S., & Sueyoshi, T. (2009). DEA as a tool for bankruptcy assessment: A comparative study with logistic regression technique. *European Journal of Operational Research*, *193*(2), 412-424. https://doi.org/10.1016/j.ejor.2007.11.036

Premachandra, I. M., Chen, Y., & Watson, J. (2011). DEA as a tool for predicting corporate failure and success: A case of bankruptcy assessment. *Omega*, *39*(6), 620-626. https://doi.org/10.1016/j.omega.2011.01.002

Rodano, G., Serrano-Velarde, N., & Tarantino, E. (2016). Bankruptcy law and bank financing. *Journal of Financial Economics*, *120*(2), 363-382. https://doi.org/10.1016/j.jfineco.2016.01.016

Strömberg, P. (2000). Conflicts of interest and market illiquidity in bankruptcy auctions: Theory and tests. *Journal of Finance*, *55*(6), 2641-2692. https://doi.org/10.1111/0022-1082.00302

Taffler, R. J. (1984). Empirical models for the monitoring of UK corporations. *Journal of Banking and Finance*, *8*(2), 199-227. https://doi.org/10.1016/0378-4266(84)90004-9

Tsai, C. F., Hsu, Y. F., & Yen, D. C. (2014). A comparative study of classifier ensembles for bankruptcy prediction. *Applied Soft Computing Journal*, *24*, 977-984. https://doi.org/10.1016/j.asoc.2014.08.047

Tseng, F., & Hu, Y. (2010). Comparing four bankruptcy prediction models: Logit, quadratic interval logit, neural and fuzzy neural networks. *Expert Systems With Applications*, *37*(3), 1846-1853. https://doi.org/10.1016/j.eswa.2009.07.081

Vapnik, V. N. (2000). *The nature of statistical learning theory* (2a ed.). New York, NY: Springer-Verlag. https://doi.org/10.1007/978-1-4757-3264-1

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3-28. https://doi.org/10.1257/jep.28.2.3

Varian, H. R. (2016). Intelligent technology. *Finance & Development*, *53*(3), 6-9.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228-1242. https://doi.org/10.1080/01621459.2017.1319839

Yang, Z., You, W., & Ji, G. (2011). Using partial least squares and support vector machines for bankruptcy prediction. *Expert Systems With Applications*, *38*(7), 8336-8342. https://doi.org/10.1016/j.eswa.2011.01.021

Yaohao, P., & Albuquerque, P. H. M. (2019). Non-linear interactions and exchange rate prediction: Empirical evidence using support vector regression. *Applied Mathematical Finance*, *26*(1), 69-100. https://doi.org/10.1080/1350486X.2019.1593866

Yaohao, P., Albuquerque, P. H. M., Camboim de Sá, J. M., Padula, A. J. A., & Montenegro, M. R. (2018). The best of two worlds: Forecasting high frequency volatility for cryptocurrencies and traditional currencies with support vector regression. *Expert Systems with Applications*, *97*, 177-192. https://doi.org/10.1016/j.eswa.2017.12.004

Zhang, M., & Zhou, Z. (2007). ML-KNN : A lazy learning approach to multi-label learning. *Pattern Recognition*, *40*(7), 2038-2048. https://doi.org/10.1016/j.patcog.2006.12.019

Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research*, *22*, 83-86. https://doi.org/10.2307/2490860