



**idp**

23.11.2021

**Nº 01**

**REVISTA**

# **DEBATES EM ECONOMIA APLICADA**

**O PODER PREDITIVO DOS MODELOS COM  
APRENDIZADO DE MÁQUINA É SUPERIOR AOS  
MODELOS TRADICIONAIS PARA ANÁLISE DO RISCO DE  
CRÉDITO?**

**ALEX CERQUEIRA PINTO**

**O PODER PREDITIVO DOS  
MODELOS COM APRENDIZADO  
DE MÁQUINA É SUPERIOR AOS  
MODELOS TRADICIONAIS PARA  
ANÁLISE DO RISCO DE CRÉDITO?**

**ALEX CERQUEIRA PINTO<sup>1</sup>**

---

<sup>1</sup> Alex Cerqueira Pinto é Mestre em Economia pelo Instituto Brasileiro de Ensino, Desenvolvimento e Pesquisa (IDP). E-mail: alexcerca10@gmail.com.



**IDP**

# DEBATES EM ECONOMIA APLICADA

[www.idp.edu.br](http://www.idp.edu.br)

O IDP é um centro de excelência no ensino, na pesquisa e na extensão nas áreas da Administração Pública, Direito e Economia. O Instituto tem como um de seus objetivos centrais a profusão e difusão do conhecimento de assuntos estratégicos nas áreas em que atua, constituindo-se um *think tank* independente que visa contribuir para as transformações sociais, políticas e econômicas do Brasil.

## DIREÇÃO E COORDENAÇÃO

### Diretor Geral

Francisco Schertel

### Coordenador do Mestrado em Administração Pública

Caio Cordeiro de Resende

### Coordenador do Mestrado em Economia

José Luiz Rossi

## CONSELHO EDITORIAL

### Coordenação

Thiago Costa Monteiro Caldeira

### Supervisão e Revisão

Renan Holtermann, Matheus Gonçalves,  
Mathias Tessmann, Milton Sobrinho e  
Alessandro Freire

### Comunicação e Marketing

Antonio Zaninetti e Daniel Jordão

### Projeto gráfico e diagramação

Juliana Vasconcelos

**Revista Técnica voltada** à divulgação de resultados preliminares de estudos e pesquisas aplicados em desenvolvimento por professores, pesquisadores e estudantes de pós-graduação com o objetivo de estimular a produção e a discussão de conhecimentos técnicos relevantes na área de Economia.

**Convidamos a comunidade** acadêmica e profissional a enviar comentários e críticas aos autores, visando o aprimoramento dos trabalhos para futura publicação. Por seu propósito se concentrar na recepção de comentários e críticas, a Revista Debates em Economia Aplicada não possui ISSN e não fere o ineditismo dos trabalhos divulgados.

**As publicações** da Revista estão disponíveis para acesso e download gratuito no formato PDF. Acesse: [www.idp.edu.br](http://www.idp.edu.br)

**As opiniões emitidas** nesta publicação são de exclusiva e inteira responsabilidade dos autores, não exprimindo, necessariamente, o ponto de vista do IDP.

**Qualquer citação** aos trabalhos da Série só é permitida mediante autorização expressa do(s) autor(es).

# SUMÁRIO

1. Introdução	6
2. Revisão da Literatura	9
3. Metodologia	13
4. Resultados	21
5. Conclusão	34
Referências	36

**Resumo:** O objetivo desse trabalho foi desenvolver e analisar, para o risco de crédito, se modelos com uso de aprendizado de máquina apresentam melhor poder preditivo comparado aos tradicionais e aplicar técnicas de interpretabilidade ao de melhor performance. A metodologia adotada corresponde a pesquisa empírica econométrica com o uso das técnicas de aprendizado supervisionado. O público-alvo foram empresas do segmento atacado. Para as variáveis do modelo foram utilizados indicadores econômicos e financeiros, retirados das demonstrações contábeis das empresas, e variáveis macroeconômicas. Os resultados indicam que o modelo de melhor capacidade preditiva foi o *XGBoost*, com curva ROC de 0.99 e acurácia de 0.98, na base teste. Na análise de interpretabilidade, via *sharp value*, os resultados corroboram o sentido econômico das variáveis. Do mesmo modo, a interpretabilidade via interações mostrou a influência da interação entre variáveis para melhora preditiva do modelo. Estes resultados corroboram a tendência de crescimento do uso de modelos com técnicas de *machine learning* na área econômica.

**Palavras-chave:** Risco de Crédito. Aprendizado de Máquina. Indicadores Financeiros. Modelo.

**Abstract:** The objective of this work was to develop and analyze, for credit risk, if models using machine learning have better predictive power compared to traditional ones and to apply interpretability techniques to those with better performance. The methodology adopted corresponds to empirical econometric research using supervised learning techniques. The target was companies from the wholesale segment. For the model variables, economic and financial indicators taken from the financial statements of the companies and macroeconomic variables were used. The results indicate that the model with the best predictive capacity was *XGBoost*, with a ROC curve of 0.99 and an accuracy of 0.98, in the test base. In the interpretability analysis, via sharp value, the results corroborate the economic meaning of the variables. Likewise, the interpretability via interactions showed the influence of the interaction between variables for the predictive improvement of the model. These results corroborate the growing trend in the use of models with machine learning techniques in the economic area.

**Key words:** Credit risk. Machine Learning. Financial indicators. Model.

## 1. INTRODUÇÃO

A principal contribuição para a perenidade de uma Instituição Financeira é a avaliação de suas atividades sob a perspectiva de desempenho e eficiência. Um sistema bancário desenvolvido e com funcionamento eficiente facilita o desenvolvimento de outras esferas de negócios na economia nacional e, portanto, influencia o desenvolvimento de todo o país. (Grmanova & Ivanova, 2018)

Todas as instituições financeiras e empresas dos mais diferentes ramos, estão sujeitos a uma pluralidade de riscos durante o ciclo operacional de seu negócio. Conhecer estes riscos é fundamental, bem como gerenciar aqueles aos quais se manifestam mais relevantes quanto a sua forma de atuação, nicho de mercado ou escala de negócio.

Damodaran (2010) descreve que o risco é onipresente em quase todas as atividades humanas e não há unanimidade acerca de uma definição para o termo. Assim, a discussão deste tema baseia-se na distinção entre o risco passível de ser quantificado de forma objetiva e o risco subjetivo.

O mercado de crédito apresenta crescente competição entre as instituições financeiras tradicionais e os novos players que estão entrando neste mercado, no qual se destacam as *fintechs* de crédito. Desta maneira, a gestão do risco de crédito, principalmente a mensuração e avaliação, são fundamentais para o segmento bancário em múltiplas concepções.

No Brasil, o Banco Central define formalmente o risco de crédito como:

“a possibilidade de ocorrência de perdas associadas ao não cumprimento pelo tomador ou contraparte de suas respectivas obrigações financeiras nos termos pactuados, à desvalorização de contrato de crédito decorrente da deterioração na classificação de risco do tomador, à redução de ganhos ou remunerações, às vantagens concedidas na renegociação e aos custos de recuperação.”  
(Bacen, 2009)

Para os autores Hull (2012) e Pesaran, Schuermann, Treutler e Weiner (2006), o risco de crédito é o principal risco enfrentado pelas instituições financeiras e, de forma geral, é objeto de rigorosa supervisão pelos reguladores nacionais, com maior necessidade de

capital para mitigação de perdas não esperadas conforme orienta o *Bank of International Settlements* – BIS (Banco de Compensações Internacionais) nos chamados Acordos de Basileia.

Por apresentar, em sua natureza, alta escalabilidade, os métodos de aprendizado de máquina é um método de maior flexibilidade em relação às técnicas econômicas de previsão mais tradicionais. Essa característica traz melhor aproximação dos dados para mensuração dos prêmios de risco. Seu uso para finanças, prevê que estes tipos de modelos, devem ter como premissa desempenho estável, e ao mesmo tempo explicativo, das proposições que levam ao descumprimento. (Gu, Kelly, & Xiu, 2018)

Desta forma, o uso pelas instituições financeiras de modelos cada vez mais robustos para predição da inadimplência é uma necessidade crescente visto: i) crescente competição no setor devido a novos atores neste mercado, conhecidos como *fintechs*; ii) redução da taxa básica de juros no Brasil e no cenário global, que reduz os ganhos de tesouraria e pressionam o *spread* de crédito; iii) forte impacto nos balanços dos bancos das provisões para créditos de liquidação duvidosa (PCLD) e das perdas com inadimplência durante os ciclos econômicos.

Assim, uma das principais atribuições dos agentes que atuam com risco de crédito constitui em desenvolver modelos para prever a probabilidade de inadimplência de um indivíduo ou empresa. Embora a regressão logística ainda seja amplamente utilizada na academia e nas áreas de risco das instituições financeiras, o processo de mensuração e identificação de riscos em si necessita de aprimoramento e desenvolvimento contínuo, uma vez que as características e fatores de risco dos agentes tomadores tendem a evoluir conforme se alteram as condições financeiras e macroeconômicas.

Deste modo, este artigo apresenta como objetivo geral descobrir se o nível de desempenho das técnicas de aprendizado de máquina para previsão de descumprimento (inadimplência) de empresas brasileiras de capital aberto é melhor do que as técnicas estatísticas tradicionais, em especial contra o de regressão logística.

Como objetivos específicos, esse estudo se propõe a: i) comparar qual modelo econométrico e de ML apresenta maior performance na predição da probabilidade de descumprimento de empresas de capital aberto, ii) identificar os indicadores contábeis e/ou de mercado mais importantes para a diferenciação de empresas segundo sua situação financeira, iii) buscar validação por meio da interpretabilidade do modelo de melhor performance com o uso das técnicas a) *sharpe value*; b) Importâncias das variáveis; e c) Medidas das Interações (*Measure Interactions*).

Para alcance dos objetivos ora propostos, foram desenvolvidos modelos de previsão não paramétricos e não lineares de risco de crédito com uso de diferentes técnicas de aprendizado de máquina (*machine learning*) e comparado os indicadores de performance dos modelos, em especial a curva ROC, para a escolha daquele com maior poder preditivo quanto a probabilidade de descumprimento (*default*). Para realizar interpretabilidade, utilizou-se o pacote SHAP em linguagem Python e o pacote IML - *Interpretable Machine Learning* na linguagem R.

Este trabalho teve seu foco de atuação circunscrito ao público de empresas segmento atacado. Outro público, como pessoas físicas, ou micro e pequenas empresas, estão fora do escopo de criação e utilização do modelo, bem como não se relacionam com as conclusões a serem apresentadas.

Os resultados apontaram que os modelos com uso de técnicas de aprendizado de máquina são ligeiramente superiores aos tradicionais quanto a predição da probabilidade de descumprimento das empresas do segmento atacado. Destaca-se, os modelos do tipo ensemble, que são modelos que combinam as decisões de outros modelos para melhor precisão da classificação final, como o *Random Forest* e o *XgBoost*, este último considerado o modelo com melhor performance dentre os 8 modelos treinados, com AUC de 0.99.

Da mesma forma, a análise de interpretabilidade observado pelo uso do *sharpe value* (valor justo) apresentou coerência e sentido econômico das variáveis de maior significância para o *output* do modelo. Verificou-se também, seguindo a análise de interpretabilidade, que as interações entre as diferentes variáveis independentes atuam de forma a elevar a capacidade preditiva do modelo e também apresentam sentido econômico.

Trabalhos deste tipo são recentes e ainda escassos com o uso das referidas técnicas, principalmente no que se refere a interpretabilidade destes modelos de *machine learning*, conhecidos como *black-box*, pois não possuem direta explicação do porquê da classificação indicada. Destaca-se também, que no ambiente regulatório das instituições financeiras, desenvolver e entender estes modelos tornou-se de grande interesse as instituições financeiras e órgãos reguladores, devido ao avanço e uso inevitável destes modelos frente aos estatísticos tradicionais.



## 2. REVISÃO DE LITERATURA

### 2.1 Modelo Teórico

Dessa forma, com o objetivo de gerir e mensurar o risco de crédito, as instituições financeiras utilizam-se dos seguintes métodos tradicionais em contraposição com as “novas” técnicas de aprendizado supervisionado. Dentre as técnicas tradicionalmente utilizadas, destaca-se na literatura o Modelo de Merton (1974).

Merton (1974) apresentou um modelo para avaliação e precificação de títulos privados por meio da relação existente entre a estrutura financeira da empresa com sua probabilidade de inadimplência. Desta forma, o autor, utilizando o modelo de Black e Scholes (1973) sobre a teoria de precificação de opções, muito utilizada pelo mercado financeiro até hoje, Merton estendeu sua aplicação para dívidas e empréstimos em geral. Desta forma, o modelo de Merton permite obter a probabilidade de *default* da firma, e, no mesmo contexto, o *spread* de crédito implícito. Conforme descrito por Souza e Corrar (2010), matematicamente temos:

$$E_0 = A_0 \cdot N(d_1) - D \cdot e^{-rT} \cdot N(d_2) \quad (1)$$

onde:

$E_0$  = Patrimônio Líquido de Mercado ( $PL_m$ ) no instante 0;

$A_0$  = Valor de Mercado do Ativo ( $A_M$ ) no instante 0;

$N(d_1)$  = Distribuição Normal Acumulada até o ponto  $d_1$ ;

$D$  = Valor de Face da Dívida;

$r$  = Taxa de juros livre de risco;

$T$  = *Duration* da Dívida, ou Prazo do CDS;

$N(d_2)$  = Distribuição Normal Acumulada até o ponto  $d_2$ ;

Neste sentido, os valores de  $d_1$  e  $d_2$  são:

$$d_1 = \frac{\ln\left(\frac{A_0}{D}\right) + (r + 0,5\sigma_A^2)T}{\sigma_A\sqrt{T}} \quad d_2 = \frac{\ln\left(\frac{A_0}{D}\right) + (r - 0,5\sigma_A^2)T}{\sigma_A\sqrt{T}}$$

Outro modelo muito utilizado pelas instituições financeiras e bureau de crédito é o modelo com o uso da técnica de Regressão Logística também conhecido como análise *logit*. Este modelo trata de uma técnica de análise multivariada, apropriada para as situações nas quais a variável dependente é categórica e assume um entre dois resultados possíveis, utilizando marcação binária de Zero (solvente) e Um (insolvente). O objetivo da regressão logística é gerar uma função matemática cuja resposta permite estabelecer a probabilidade, de uma observação pertencer a um grupo previamente determinado, em razão do conjunto de variáveis independentes. Deste modo, os coeficientes estimados pelo modelo de regressão indicam a importância de cada variável independente para a ocorrência do evento. (Brito & Assaf Neto, 2009)

Matematicamente a regressão logística é descrita como:

$$\text{logit}[\theta(x)] = \log \left[ \frac{\theta_x}{1-\theta_x} \right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n \quad (2)$$

onde  $\alpha$  é constante da equação e  $\beta$  são os coeficientes das variáveis predictoras.

## 2.2 Literatura Nacional e Internacional

As literaturas nacionais e internacionais apresentam uma infinidade de estudos no que diz respeito à análise da do risco de crédito e diferentes métodos para sua avaliação e mensuração.

Neste sentido, Brito e Assaf Neto (2009) construíram um modelo tradicional de risco de crédito, com usa da técnica de regressão logística, onde utilizaram amostra de 60 empresas de capital aberto no período entre 1994 e 2004, classificando-as como solventes ou insolventes a partir do pedido de concordata registrado nos relatórios Boletim Diário de Informações, publicados pela Bovespa e cadastro de companhias abertas da Comissão de Valores Mobiliários (CVM). As variáveis independentes utilizadas foram índices financeiros calculados a partir das demonstrações contábeis das empresas do penúltimo exercício anterior ao ano do evento de *default*. O modelo final apresentou excelente poder preditivo, com curva ROC em 0.97, e foi composto pelo intercepto e quatro variáveis explicativas, a saber: i) lucros retidos sobre ativo; ii) endividamento financeiro; iii) capital de giro líquido; e iv) saldo de tesouraria sobre vendas.

Soares e Rebouça (2015), apresentaram trabalho para cálculo de risco de crédito que testou vários modelos de previsão de insolvência de empresas brasileiras de capital aberto, onde utilizaram-se de amostra com 21 empresas insolventes e 66 empresas solventes, escolhidas conforme a distribuição setorial do primeiro grupo. Utilizaram as técnicas: análise discriminante, regressão logística, árvores de classificação e regressão (CART) e redes neurais artificiais (ANN), este último com desempenho consideravelmente superior aos demais.

Guimarães e Moreira (2008) propõem um modelo de previsão de insolvência baseado em indicadores contábeis com o uso da análise discriminante. Os autores utilizaram amostra composta por informações financeiras e contábeis de 116 empresas de capital aberto de 17 setores diferentes, entre 1994 e 2003, coletadas do banco de dados do Sistema de Análise de Balanços de Empresas do IBMEC. Os indicadores contábeis das empresas com *default* foram extraídos dos demonstrativos contábeis referentes a um ano antes da entrada no estado de insolvência. Assim, com relação às variáveis preditoras do modelo, os autores confirmaram o poder discriminatório daquelas que evidenciam as decisões financeiras sobre estrutura de ativos, estrutura de capital e geração de caixa.

Também no contexto do risco de crédito, Jackson e Wood (2013) propõem e testam modelos para previsão de insolvência para falência de empresas baseados em dados contábeis e investigam sua eficácia. Assim, eles utilizaram os dados do Reino Unido, com as empresas que faliram entre 2000 e 2009 considerando como população todas as empresas não financeiras listadas na London Stock Exchange (LSE), em um conjunto de 101 empresas com falência. Dos treze modelos testados, os quatro modelos com melhor desempenho são modelos de reivindicações contingentes com base em opções de compra e barreira europeias. Outra conclusão, baseia-se que as variáveis de fluxo de caixa e dívida total, apresentam boa capacidade preditiva.

Em pesquisa semelhante, Luo, Wu, e Wu (2017) aplicam sobre a perspectiva de classificação de risco de crédito, modelagem com o uso de algoritmos de *Deep Learnig* para rating de crédito de empresas globais, que atuam em diversos setores, após a crise de 2008, usando como variável o *Credit Defaul Swap* (CDS) das empresas. Os resultados do teste indicam que o *Deep Learnig* supera outros algoritmos e podem facilmente classificar o risco de crédito antecipadamente das empresas.

Por fim, Xia, Liu, Li, e Liu (2017) propuseram um modelo de pontuação de risco de crédito usando como base no modelo de *Extreme Gradient Boosting*, conhecido como *XGBoost*. O modelo compreende principalmente três etapas que foram utilizadas para calibrar o modelo no objetivo de oferecer assertivo poder classificatório. Os resultados do

trabalho demonstram que o modelo proposto pelos autores supera, em média, os modelos tradicionais em quatro medidas de performance: precisão, taxa de erro, área sob a curva (AUC), além de melhor abrangência e interpretabilidade na pontuação de crédito.

Assim, a tabela 1 apresenta o resumo dos autores citados, as técnicas utilizadas, público-alvo e principais conclusões:

**Tabela 1 – Resumo dos autores apresentados na Revisão Bibliográfica**

AUTORES	TÉCNICA UTILIZADA	PÚBLICO-ALVO	DADOS	CONCLUSÕES
<b>Brito e Assaf Neto (2009)</b>	Regressão Logística	60 empresas de capital aberto no período entre 1994 e 2004.	Índices financeiros calculados a partir das demonstrações contábeis do penúltimo exercício anterior ao ano default.	Curva ROC de 0,978 e quatro variáveis explicativas, a saber: (i) lucros retidos sobre ativo; (ii) endividamento financeiro; (iii) capital de giro líquido; e (iv) saldo de tesouraria sobre vendas.
<b>Soares e Rebouça (2015)</b>	Análise discriminante, Regressão logística, Árvores de classificação e regressão e Redes Neurais Artificiais	Amostra com 21 empresas insolventes e 66 empresas solventes.	Informações contábeis referentes ao ano imediatamente anterior à ocorrência de <i>default</i> durante os anos de 2002 a 2012.	A Melhor performance foi com a técnica Redes Neurais. Variáveis mais importantes foram: a relação entre Patrimônio Líquido e Capitais de Terceiros; e LAJIR. A decisão de tomar dívida é o principal para o risco de insolvência.
<b>Guimarães e Moreira (2008)</b>	Análise Discriminante	Amostra composta por 116 empresas de capital aberto entre 1994 e 2003.	Indicadores financeiros e contábeis	Poder discriminatório daquelas que evidenciam as decisões financeiras sobre estrutura de ativos, estrutura de capital e geração de caixa. Foi considerado relevante 4 indicadores financeiros.

<b>Jackson e Wood (2013)</b>	Total de 13 modelos de <i>machine learning</i> testados.	Empresas não financeira do Reino Unido listadas na London Stock Exchange (LSE)	Testam modelos para previsão de insolvência para falência de empresas baseados em dados contábeis.	Conclusão baseia-se que as variáveis de fluxo de caixa e dívida total, são determinantes na capacidade preditiva da empresa.
<b>Luo et al. (2017)</b>	<i>Deep Learnig</i> , Regressão Logística Multinomial (MLR), <i>Perceptron Multilayer</i> (MLP) e SVM.	Rating de crédito de empresas globais, que atuam em diversos setores, após a crise de 2008.	Usaram como variável o <i>Credit Defaul Swap</i> (CDS) das empresas.	Os resultados indicam que o uso de <i>Deep Learning</i> supera outros algoritmos e podem facilmente classificar o risco de crédito antecipadamente das empresas.
<b>Xia et al. (2017)</b>	<i>XGBoost</i>	Mercado PF e PJ.	São utilizados dados de crédito da Alemanha, Áustria e Taiwan do repositório da UCI e de plataformas P2P de empréstimos.	O modelo proposto pelos autores supera modelos tradicionais em quatro medidas de performance: precisão, taxa de erro, área sob a curva (AUC).

Fonte: Elaboração própria.

## 3. METODOLOGIA

### 3.1 Metodologia de Pesquisa

O presente artigo busca realizar uma pesquisa empírica econométrica com o desenvolvimento de um modelo de risco, com uso de ML. Deste modo, a criação de um modelo possui diversas etapas que compreendem: formulação dos objetivos e hipóteses do modelo, coleta e tratamento dos dados, criação e marcação do *target* do modelo, o pré-processamento dos dados, hiperparametrização, treino, teste e validação do modelo.

O indicador de comparação dos modelos para escolha da melhor técnica será feito por meio da área abaixo da curva ROC e Matriz de Confusão.

Por se tratar da criação e comparação de modelos com técnicas de *machine learning*, será utilizada a metodologia *Cross Industry Standard Process for Data Mining* (Crisp-DM). Esta metodologia reúne as algumas das melhores práticas mineração de dados, de forma que o processo de tratamento de dados e modelagem seja o mais produtivo e eficiente possível. (Tukey, 1977).

Neste sentido, a forma de aprendizado dos algoritmos de *machine learning* pode ser classificada em: aprendizado supervisionado, não supervisionado e aprendizado por reforço. Este trabalho utilizou o aprendizado supervisionado.

No que se refere ao aprendizado supervisionado, é fornecido um conjunto de dados rotulados onde é para cada observação é informado a saída correta ou sua categoria. No modelo de risco, para as características extraídas de cada variável da amostra indica-se ao modelo se a empresa apresenta descumprimento ou não. Em síntese, modelos supervisionados apresentam a variável dependente, também chamada de *target*, nos dados de treinamento do modelo, assim o algoritmo busca reconhecer um padrão de comportamento das observações para posteriormente realizar a classificação.

Durante as etapas de treino e validação do modelo, será utilizada a metodologia de validação cruzada com o objetivo de: i) utilizar os melhores hiperparâmetros possíveis, ii) evitar sobreajustamento (*overfitting*) e; iii) elevar os indicadores de performance com maior variabilidade dos dados de treino e teste.

Os dados utilizados no artigo são os disponibilizados pela Comissão de Valores Mobiliários – CVM, através dos *datasets* para *download* ou por intermédio de pacote GetDFPData disponível para a linguagem R. Os dados são referentes aos balanços reportados pelas empresas no período de 2011 a 2019. Este pacote fornece uma interface aberta para todas as demonstrações financeiras distribuídas pela B3 e pela CVM nos sistemas DFP, FRE e FCA. O instrumento não só permite o *download* dos dados mas também consolida as diferentes notações contábeis, ajusta à inflação e torna as tabelas disponíveis para pesquisa dentro de um formato tabular. (Perlin, 2017)

Assim, como realizado por Brito e Assaf Neto (2009), a população utilizada no estudo, a partir da qual a amostra será selecionada, compreende as empresas do segmento atacado não financeiras registradas na CVM. A retirada de empresas financeiras foi devido as mesmas possuírem estrutura patrimonial, distribuição e características de ativos e passivos,

diferente das demais empresas não financeiras. Assim, foram colhidas 8.766 observações trimestrais de um total de 473 empresas do período analisado. Os dados foram organizados em formato painel empilhados, ou seja, não considerando seu aspecto temporal.

A escolha de empresas dessa natureza e porte ocorre por estas serem empresas com informações de domínio público e facilidade de divulgação de informações contábeis. Os dados coletados na forma de informações contábeis dessas empresas, foram transformados em indicadores econômicos/financeiros para serem utilizados como variáveis preditivas para os modelos a serem testados.

A relação de variáveis independentes testadas nos modelos estão descritas nas Tabelas 2 e 3. Trata-se de 27 variáveis construídas a partir de índices contábeis das empresas e índices e indicadores macroeconômicas.

**Tabela 2 – Relação de Índices Contábeis para uso como variáveis dos modelos**

Código da Variável	Índice Financeiro	Fórmula do Índice
V1	Liquidez geral	$(\text{Ativo Circulante} + \text{Realizável a Longo Prazo}) / (\text{Passivo Circulante} + \text{Exigível a Longo Prazo})$
V2	Liquidez corrente	$\text{Ativo Circulante} / \text{Passivo Circulante}$
V3	Liquidez seca	$(\text{Ativo Circulante} - \text{Estoques}) / \text{Passivo Circulante}$
V4	Liquidez imediata	$\text{Disponível} / \text{Passivo Circulante}$
V5	Retorno sobre o patrimônio líquido	$\text{Lucro Líquido} / \text{Patrimônio Líquido inicial}$
V6	Retorno sobre o ativo	$\text{LAJIR} / \text{Ativo Total}$
V7	Retorno sobre vendas	$\text{Lucro Líquido} / \text{Vendas Líquidas}$
V8	Giro do ativo	$\text{Vendas Líquidas} / \text{Ativo Total}$
V9	Margem operacional	$\text{LAJIR} / \text{Vendas Líquidas}$
V10	Lucro operacional sobre despesas financeiras	$\text{LAJIR} / \text{DF}$
V11	Patrimônio líquido sobre ativo	$\text{Patrimônio Líquido} / \text{Ativo Total}$

V12	Lucros retidos sobre ativo	$(\text{Lucros Acumulados} + \text{Reserva de Lucros}) / \text{Ativo Total}$
V13	Patrimônio líquido sobre exigível total	$\text{Patrimônio Líquido} / (\text{Passivo Circulante} + \text{Exigível a longo Prazo})$
V14	Endividamento total	$(\text{Passivo Circulante} + \text{Exigível a longo Prazo}) / \text{Ativo Total}$
V15	Endividamento de curto prazo	$\text{Passivo Circulante} / \text{Ativo Total}$
V16	Endividamento financeiro	$(\text{Passivo Circulante Financeiro} + \text{Exigível a longo Prazo Financeiro}) / \text{Ativo Total}$
V17	Imobilização do patrimônio líquido	$\text{Ativo Permanente} / \text{Patrimônio Líquido}$
V18	Estoques sobre ativo	$\text{Estoques} / \text{Ativo Total}$
V19	Capital de giro líquido	$(\text{Ativo Circulante} - \text{Passivo Circulante}) / \text{Ativo Total}$
V20	Necessidade de capital de giro	$(\text{Ativo Circulante Operacional} - \text{Passivo Circulante Operacional}) / \text{Ativo Total}$
V21	Saldo de tesouraria sobre ativo	$(\text{Ativo Circulante Financeiro} - \text{Passivo Circulante Financeiro}) / \text{Ativo Total}$
V22	Saldo de tesouraria sobre vendas	$(\text{Ativo Circulante Financeiro} - \text{Passivo Circulante Financeiro}) / \text{Vendas Líquidas}$
V23	Fluxo de caixa operacional sobre ativo	$\text{Fluxo de Caixa das Operações} / \text{Ativo Total}$
V24	Fluxo de caixa operacional sobre exigível total	$\text{Fluxo de Caixa das Operações} / (\text{Passivo Circulante} + \text{Exigível a longo Prazo})$
V25	Fluxo de caixa operacional sobre endividamento financeiro	$\text{Fluxo de Caixa das Operações} / (\text{Passivo Circulante Financeiro} + \text{Exigível a longo Prazo Financeiro})$

Fonte: Brito, Assaf Neto, e Corrar (2009). Adaptada. Elaboração própria.



**Tabela 3 – Relação de Variáveis Macroeconômicas**

Código da Variável	Índice Financeiro	Tipo
M1	Taxa Selic Over a.t.	Numérica
M2	Desvio PIB Potencial - % a.a.	Numérica
M3	Risco Brasil (Credit default swap)	Numérica
M4	Índice de Utilização da Capacidade instalada -UCI	Numérica
M5	Resultado Primário do Governo Geral	Numérica
M6	VIX - Índice Volatilidade	Numérica
M7	Diferencial de juros entre Brasil e EUA	Numérica
M8	Diferencial de juros entre Brasil e Países Emergentes	Numérica
M9	Rentabilidade IBOVESPA	Numérica
M10	Rentabilidade NTN-B - Prêmio	Numérica
M11	PIB - Variação Trimestral - %	Numérica
M12	PIB - Variação Média Móvel 2 anos - %	Numérica
M13	PIB - Variação Média Móvel 3 anos - %	Numérica
M14	Inflação - IPCA	Numérica
M15	Taxa de Câmbio - Real/Dólar - Nominal	Numérica
M16	Taxa de Câmbio - Real/Dólar - Real - Índice	Numérica

Fonte: Elaboração própria.

A marcação de descumprimento na base de dados será a variável dependente do modelo, também chamada de *target*. A identificação das empresas insolventes foi realizada por meio do relatório de situação cadastral das empresas registradas na CVM, onde são identificadas empresas em falência e recuperação judicial. Deste modo, para indicar caráter preditivo ao modelo, a marcação na base de dados para empresas em *default* foi realizada nos indicadores de balanço um ano antes da data de registro de falência ou recuperação judicial. Deste modo, a base de dados foi marcada, na sua *target* como sendo 1(um) para *default*, e 0 (zero) para não-*default*.

Deste modo, após exclusão da base de dados de observações que impediriam a criação do modelo (como ausências de informações relevantes ou *missings*) a base de dados final de modelagem utilizou dados de 423 empresas em 7734 observações trimestrais, de modo que 384 marcadas foram marcadas como *default*, representando 4,96% da amostra. Por se tratar de dados não balanceados, também foram avaliados os modelos com os dados balanceados na proporção de 1/1 por meio do algoritmo SMOTE, desenvolvido por Chawla, Bowyer, Hall, e Kegelmeyer (2002) e aplicado na linguagem R. Neste método, os autores desenvolveram metodologia de sobreamostrar a classe minoritária envolvendo a criação de exemplos sintéticos de classe minoritária.

### 3.2 Técnicas Econométricas e de *Machine Learning*

Conforme o objetivo do trabalho, serão utilizadas algumas técnicas de modelagem com o uso do chamado ferramental de *machine learning* (aprendizado de máquina). As técnicas de ML são consideradas um dos avanços mais recentes e importantes da matemática aplicada, com diversas aplicações na medicina, economia, finanças, robótica, e diversos segmentos da indústria e serviços. Praticamente toda a sociedade é, de alguma forma, será fortemente impactada pelo crescente uso destas ferramentas.

Conforme descrito por Tian, Shi, e Liu (2012), o *machine learning* é uma sequência da ciência da computação, que, em conjunto com aplicações da estatística, deu aos softwares a habilidade de aprender padrão de comportamento, sendo utilizados principalmente em problemas de classificação e predição.

Desta forma, será dada rápida contextualização das técnicas a serem utilizadas no projeto, tais quais: *Naive Bayes*; *Random Forest*; *Extreme Gradient Boosting - XGboost*; *Linear Support Vector Machin* – SVM; e Redes Neurais Artificiais.

O algoritmo *Naive Bayes* é um algoritmo de classificação binária probabilístico muito utilizado em *machine learning*. Baseado no Teorema de Bayes, este modelo é frequentemente aplicado em processamento de linguagem natural e diagnósticos médicos, entre outros. O teorema de Bayes trata sobre probabilidade condicional, isto é, a probabilidade de o evento A ocorrer, dado o evento B. (Lewis, 1998)

O algoritmo supõe que há independência entre as variáveis do modelo, ou seja, o algoritmo assume que as probabilidades são condicionalmente independentes a target ao

invés de calcular o valor das probabilidades relacionada entre cada atributo. Por isso seu nome *naive* - ingênuo. (Lewis, 1998)

Desta forma, a rede bayesiana é descrita assim:

$$P\left(C = \frac{c_k}{X} = x\right) = P(C = c_k) x \frac{P\left(X = \frac{x}{c} = c_k\right)}{P(x)} \quad (3)$$

$$P(X = x) = \sum_{k'=1}^{ec} P\left(X = \frac{x}{c} = c_{k'}\right) x P(C = c_{k'}) \quad (4)$$

onde:

$P(c|x)$ : é a probabilidade da hipótese “c” dado a observação “x”. Isso é conhecido como probabilidade posterior;

$P(x|c)$ : é a probabilidade da observação “x” dado que a hipótese “c” é verdadeira;

$P(c)$ : é a probabilidade de hipótese “h” ser verdadeira (independentemente dos dados). Isso é conhecido como probabilidade anterior de “h”;

$P(x)$ : é a probabilidade da observação “x” (independentemente da hipótese).

No universo do aprendizado de máquina, uma das formas de melhorar a capacidade dos algoritmos é por meio da combinação destes. Neste artigo, faz parte desta classe de modelos o *Random Forest* e *Xgboost*.

Estes algoritmos são conhecidos como do tipo *Ensemble*, ou seja, que combinam modelos simples e de baixo poder preditivo (*weak models*), para produzir um único forte, robusto e com maior acurácia. As principais metodologias de Ensemble são: *Bagging* e *Boosting*.

A metodologia *bagging*, proposta utilizada no *Random Forest*, foi proposta por Breiman (2001), tem por objetivo reduzir a variância das predições. Vários algoritmos são treinados separadamente em diversas reamostragens com reposição do mesmo conjunto de treinamento. De maneira geral, o método *bagging* se baseia na:

- a. Construção das bases de treinamento utilizando bootstrap na base de treinamento original. Amostragem com reposição para formação dos dados;

- b. Criar múltiplos algoritmos construídos para cada conjunto de dado reamostrado;
- c. Combinar os algoritmos: As previsões são combinadas utilizando médias, moda, mediana para regressão ou voto majoritário para problemas de classificação.

No caso da *Random Forest*, este modelo combina várias árvores de decisão e os valores combinados tendem a ser mais robusto que o valor gerado por um único modelo. O modelo constrói várias árvores pouco correlacionadas, onde a principal melhoria das árvores combinadas é a redução da variância. (James, Witten, Hastie, & Tibshirani, 2013)

Destaca-se como vantagem da técnica de *Random Forest* a capacidade de lidar com dados em grandes volumes e com muitas variáveis e a habilidade de identificar as variáveis mais significativas dentro de um conjunto de variáveis de entrada. Em contrapartida, como desvantagem, o modelo pode facilmente superajustar a base de dados de treino (*overfitting*), assim como dar maior importância para variáveis altamente categorizadas, mesmo que estas não possuam alto poder explicativo, além deste modelo ser de difícil interpretação. (James et al, 2013)

Por outro lado, segundo James et al. (2013), no método *boosting* os algoritmos são aplicados de maneira sequencial de forma que a cada iteração o algoritmo aplicado utiliza a informação do algoritmo aplicado na iteração anterior, ou seja, a cada iteração ajusta-se o algoritmo usando os resíduos do modelo (erros) da interação anterior como a variável dependente, no lugar da variável resposta.

Deste modo, o *XGboost* é um algoritmo de aprendizado de máquina do tipo *boosting*, baseado em árvore de decisão e que utiliza uma estrutura de *gradient boosting*. Este é um método que mais tem ganhado competições de *machine learning* na plataforma Kaggle do Google, muitas vezes combinado com redes neurais profundas.

O fator mais importante por trás do sucesso do *XGBoost* é sua escalabilidade em todos os cenários devido a sua otimização algorítmica. O sistema roda mais de dez vezes mais rapidamente do que as soluções populares existentes em uma única máquina e escala para bilhões de exemplos em configurações distribuídas ou com pouca memória. (Chen & Guestrin, 2016)

Outra técnica a ser testada no artigo é o *Support Vector Machine* – SVM. Desenvolvida por Boser, Guyon e Vapnik (1992), este é um algoritmo de *machine learning*, considerado de aprendizado supervisionado utilizado para classificação.

Segundo Betancourt (2005), o SVM tem como vantagens: (i) a facilidade de treinar; (ii) não apresenta um ótimo local, como nas redes neurais; (iii) escala relativamente bem para dados em espaços de alta dimensão. (iv) a relação entre a complexidade do classificador e o erro pode ser explicitamente controlado; e (v) dados não tradicionais, como caracteres, podem ser usados como entrada, em vez de vetores de recursos.

Por outro lado, a fraqueza do SVM é a necessidade de uma função "boa" do *kernel*, ou seja, são necessárias metodologias eficientes para ajustar os parâmetros de inicialização do SVM. (Betancourt, 2005)

Por fim, a Rede Neural Artificial (ANN) será a última técnica a ser utilizada no projeto. Este modelo foi descrito por McCulloch e Pitts (1943), que criaram um sistema que reproduz as características básicas de um neurônio humano, o perceptron.

Desta forma, as ANNs são uma técnica de processamento de informação inspirada pelo sistema nervoso humano. Conforme descrito por Haykin (2007), o cérebro humano pode ser considerado um sistema de processamento de informação extremamente complexo, não linear e paralelo, que realiza diversas atividades de maneira muito mais eficaz que sistemas computacionais.

## 4. RESULTADOS

### 4.1 Escolha das variáveis preditoras

A partir das variáveis coletadas, em um primeiro momento, avaliamos a relação de correlação entre as variáveis preditoras conforme apresentado a figura 1. Um ponto de corte estabelecido é o não uso de variáveis com correlação entre si maior que 0.6, optando por apenas uma das variáveis.

De forma adicional para escolha das variáveis, foi realizada avaliação via importância relativa das variáveis, através da técnica de árvore de decisão, para selecionar aquelas com maior poder preditivo aos modelos conforme destaca a figura 2.

Figura 1 - Matriz de Correlação de Pearson das variáveis predictoras

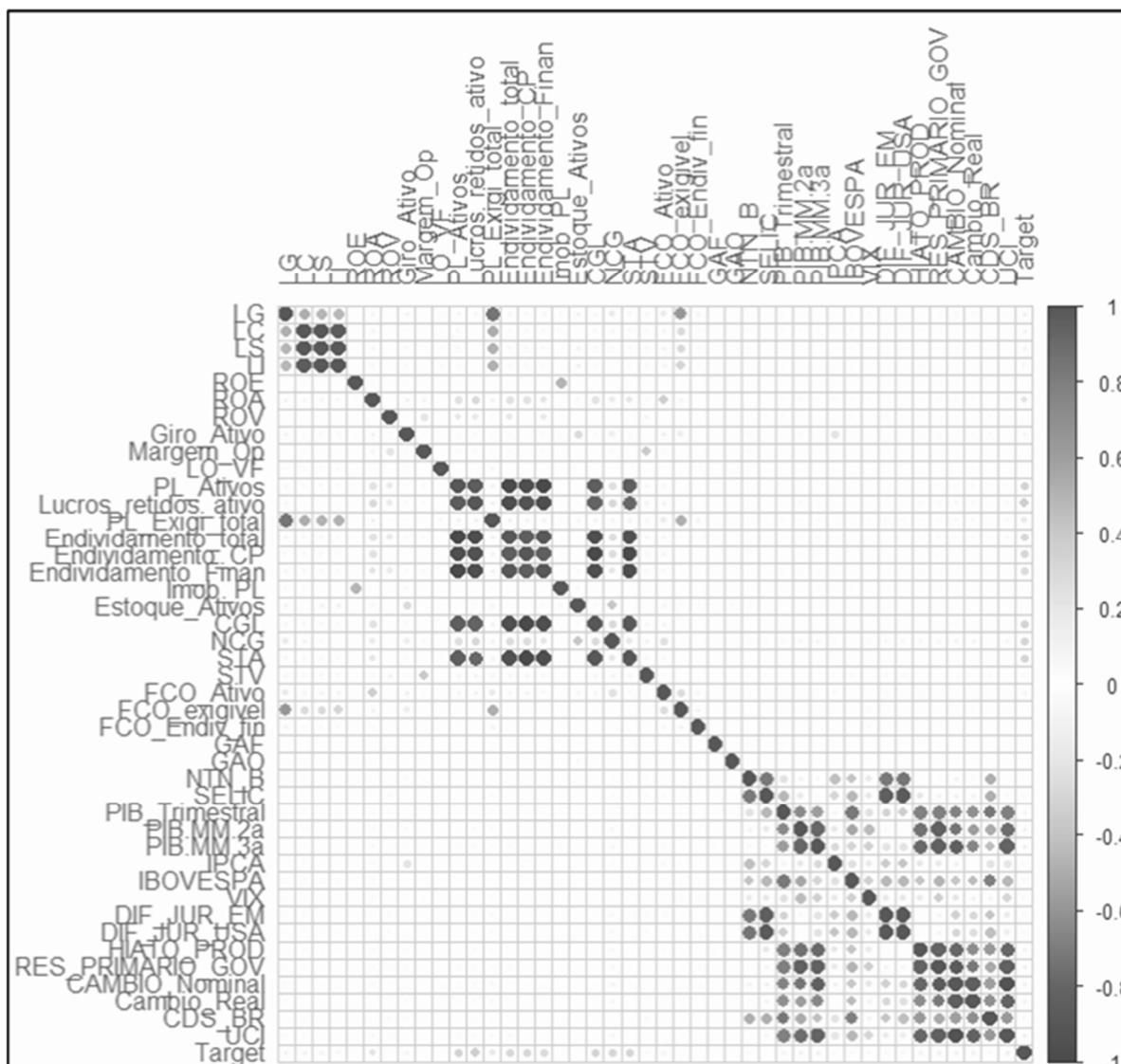
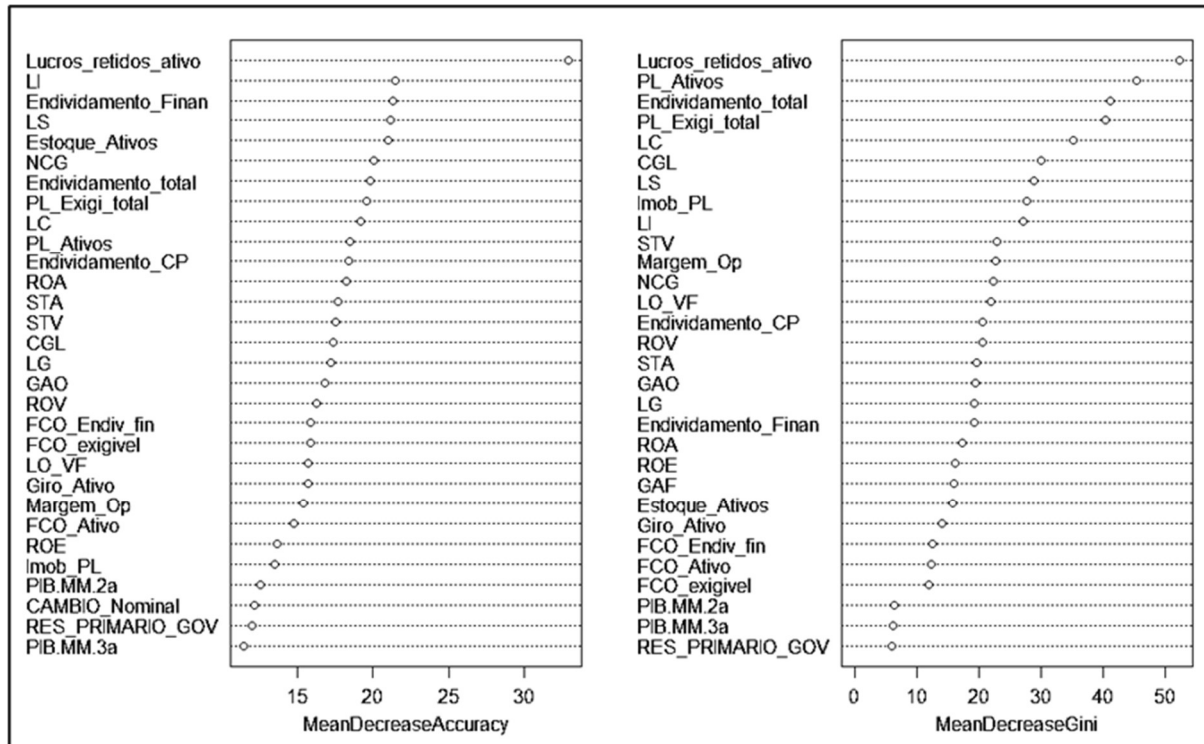


Figura apresenta a matriz de correlação de todas as variáveis inicialmente testadas no modelo versus todas as demais. Coloração escura indica alta correlação.

Fonte: Dados da Pesquisa. Elaboração própria.

Figura 2 - Importância Relativa das Variáveis



Esta figura apresenta, por ordem crescente, quais variáveis o modelo considera mais importantes para predição.

Fonte: Dados da Pesquisa. Elaboração própria.

Outra forma de seleção de variáveis utilizada foi a *Recursive feature selection* com a utilização de validação cruzada. Por meio deste método, as 5 variáveis com maior poder preditivo foram: Lucros Retidos/Ativo, Liquidez Imediata, Estoque/Ativos, Endividamento de Curto Prazo e Liquidez Seca.

Por intermédio dos parâmetros e técnicas acima citados, das 43 variáveis iniciais foram escolhidas 24 variáveis de forma a minimizar correlações e *overfitting*, maximizar o poder preditivo e manter a capacidade de generalização do modelo. Todas estas variáveis foram aplicadas e testadas em todos os modelos. A relação final de variáveis consta na tabela 4 abaixo:

**Tabela 4 – Relação de Variáveis Finais para uso nos modelos**

Variáveis Finais		
Liquidez Seca (LS)	Giro do Ativo	Grau de Alavancagem Financeira (GAF)
Saldo de Tesouraria sobre Vendas (STV)	Necessidade de Capital de Giro (NCG)	IPCA
Lucro Operacional /Despesas Financeiras	Retorno sobre Vendas (ROV)	IBOVESPA
Fluxo de Capital Operacional sobre Exigível a LP	Fluxo de Caixa Operacional sobre Endividamento Financeiro	VIX
Patrimônio Líquido sobre Exigível Total	Lucros Retidos sobre Ativos	Diferencial Juros BR/Emergentes
Estoque sobre Ativos	Imobilização sobre Patrimônio Líquido	Hiato do Produto
Retorno sobre o patrimônio líquido (ROE)	Margem Operacional	Cambio Real
Retorno sobre Ativos (ROA)	Grau de Alavancagem Operacional (GAO)	CDS - Brasil

Notas: As vinte e quatro variáveis finais apresentadas nesta tabela serão utilizadas como variáveis preditivas em todos as técnicas testadas. Esta seleção baseou-se na análise correlação das variáveis, grau de importância relativa das variáveis e seleção recursiva de variáveis.

Fonte: Elaboração própria.

## 4.2 Comparativo dos Modelos e escolha da melhor performance

Na etapa do desenvolvimento dos modelos, foi realizada a segmentação aleatorizada da base de observação na proporção de 80% para treino dos modelos e 20% para teste *out-of-time* dos modelos. Do mesmo modo, durante o treino dos modelos com o uso de 80% dos dados, foi utilizada a metodologia de treinamento com validação cruzada (*cross validation*). Assim, os resultados apresentados nessa sessão são referentes a aplicação dos modelos treinados na amostra fora do tempo (*out-of-time*).



Após a aplicação aos dados as técnicas mencionadas, utilizando todos os modelos com a melhor configuração de hiperparâmetros possível, verificou-se que os algoritmos com uso de aprendizado de máquina apresentaram poder preditivo melhor do que a tradicional regressão logística, como se pode observar nas Tabelas 5 e 6 abaixo. Do mesmo modo, o *XGBoost* foi escolhido como a técnica que apresentou a melhor performance, apresentando curva ROC de 0.99 nos dados em sua distribuição não balanceada e 0.97 com os dados balanceados.

**Tabela 5 – Indicadores de performance dos algoritmos (dados em nível)**

	Naive Bayes	SVM - Linear	SVM - Polynomial	SVM - Radial	Regressão Logística	Redes Neurais	Random Forest	XGBoost
<b>Acurácia</b>	0,9496	0,9528	0,9509	0,9651	0,9655	0,9638	0,9780	0,9838
<b>Precision</b>	0,6666	0,6538	0,9000	0,8478	0,8510	0,7258	0,9615	1,0000
<b>Recall</b>	0,0970	0,2098	0,1071	0,4534	0,3539	0,5357	0,6097	0,6600
<b>F1</b>	0,1700	0,3177	0,1914	0,5909	0,5000	0,6164	0,7462	0,7967
<b>ROC</b>	0,8900	0,9100	0,9400	0,9400	0,9400	0,9600	0,9700	0,9900

Notas: Tabela elaborada com os dados extraídos da aplicação da matriz de confusão (*confusion matrix*) em cada modelo treinado nos dados de teste fora da amostra com a aplicação de calibração de seus hiperparâmetros.

Fonte: Elaboração própria.

**Tabela 6 – Indicadores de performance dos algoritmos (dados balanceados)**

	Naive Bayes	SVM - Linear	SVM - Polynomial	SVM - Radial	Regressão Logística	Redes Neurais	Random Forest	XGBoost
<b>Acurácia</b>	0,9347	0,9244	0,9166	0,9024	0,9257	0,9011	0,9586	0,9606
<b>Precision</b>	0,4054	0,3629	0,355	0,3251	0,3856	0,3236	0,5625	0,5798
<b>Recall</b>	0,5625	0,6125	0,7500	0,8250	0,7375	0,8375	0,9000	0,8625
<b>F1</b>	0,4712	0,4558	0,4819	0,4664	0,5064	0,4669	0,6923	0,6934
<b>ROC</b>	0,8700	0,8800	0,9000	0,9100	0,9100	0,9200	0,9700	0,9700

Notas: Tabela elaborada com os dados extraídos da aplicação da matriz de confusão (*confusion matrix*) em cada modelo treinado nos dados de teste fora da amostra com a aplicação de calibração de seus hiperparâmetros.

Fonte: Elaboração própria.

## 4.3 Aplicação da interpretabilidade do modelo campeão como validação

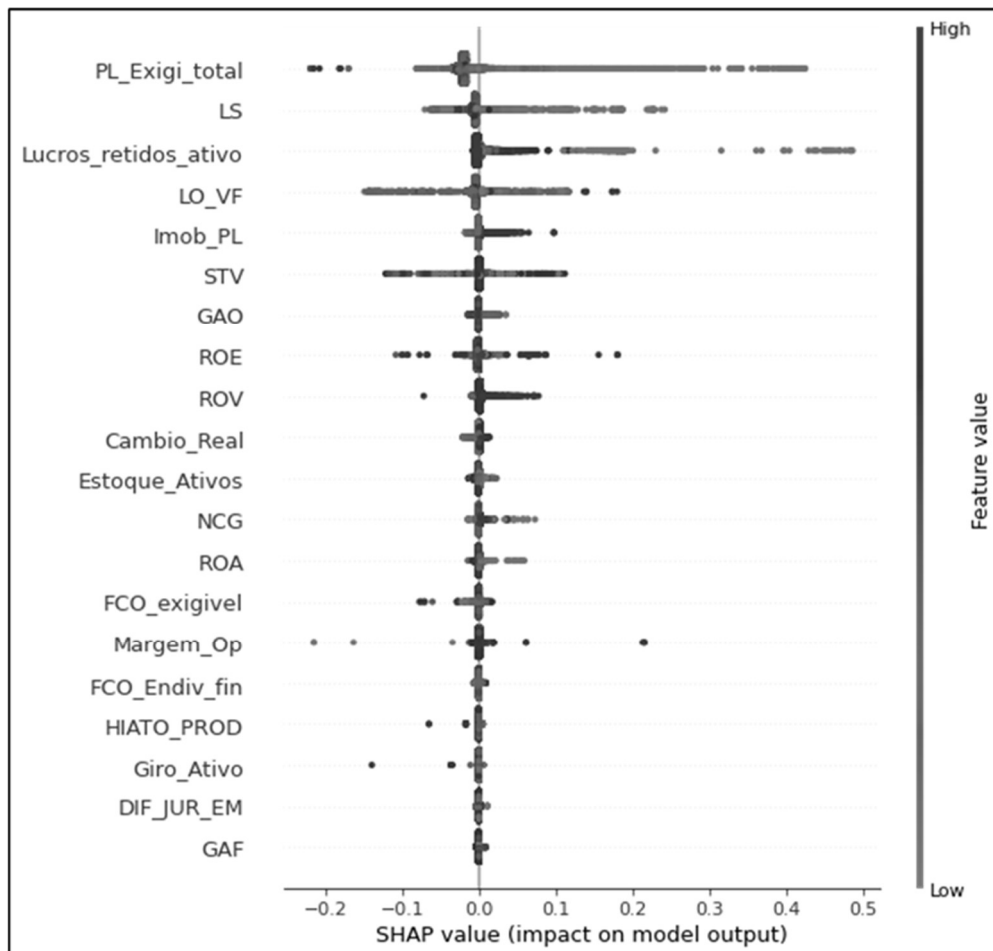
Para o modelo *XGBoost*, que apresentou a melhor capacidade preditiva, de forma a complementar este trabalho, foi realizada a análise da interpretabilidade de suas estimativas. Modelos com esse tipo de técnica são considerados do tipo “*black-box*” por se tratar de um algoritmo não tradicional onde não se pode observar os valores dos betas estimados para as variáveis, bem como realizar testes estatísticos padrões.

Os modelos de *machine learning*, conhecidos como “*black-box*”, são cada mais estudados com o objetivo de clarificar a relação entre as variáveis explicativas e o output do modelo. Importante ressaltar que a interpretabilidade do modelo visa, inclusive, avaliar a qualidade do modelo, à medida que permite a compreensão da relação entre as variáveis explicativas e o *output*.

Para os objetivos deste trabalho, serão aplicadas três ferramentas que buscam dar interpretabilidade aos modelos: a) *Shapley Value*; b) Importância e c) *Interactions*.

A partir do conceito de *Shapley Value*, oriundo da Teoria dos Jogos, Lundberg e Lee (2017) desenvolveram um método aplicado aos modelos de *machine learning* que confere interpretabilidade aos modelos. Neste contexto, o *shapley value* mede a contribuição de cada variável na construção do *output*, ou seja, o valor justo que cada variável influencia no resultado do modelo. Sendo assim, a relação entre esses valores e os valores das covariáveis permite avaliar o significado econômico de cada variável.

Figura 3 - Análise de Shap Value



A figura apresenta, por ordem crescente, as variáveis com maior valor justo, ou seja, mas mais importantes para o modelo. As observações a direita da linha central, indicam que os valores de cada variável contribuem para maior probabilidade de a empresa estar em descumprimento (*default*).

Fonte: Elaboração própria. Realizado em Python – Pacote Shap.

O gráfico da Figura 3 ordena as variáveis por importância, e corrobora os resultados apresentados pelo modelo no que se refere a interpretação econômica dos resultados. Deste modo, o *Shapley Value* indica uma relação inversa entre as variáveis representativas “PL/Exigível Total”, “Liquidez Seca” e “Lucros Retidos/Ativos” e o valor predito. Ou seja, quanto menor (*low*) os valores dessas variáveis maior a probabilidade de descumprimento.

O método acima permite ainda que seja realizada uma interpretabilidade local do modelo. Nesse caso, os resultados explicam individualmente cada valor previsto. Esse método é mais útil caso haja necessidade de explicar um valor previsto individualmente.

**Figura 4 - Importância dos Atributos**

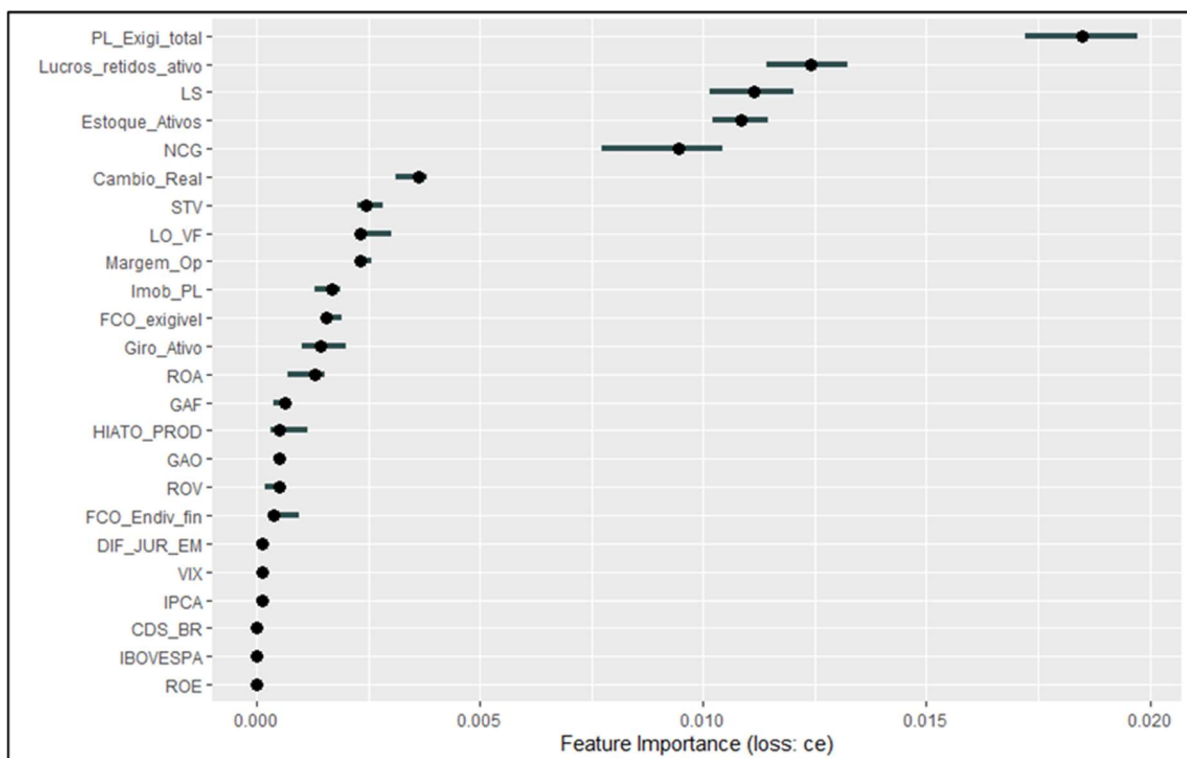


Figura apresenta, por ordem crescente, as variáveis mais relevantes para o modelo segundo a medida de erro “classification error”.

Fonte: Elaboração própria. Realizado em R – Pacote lml.

Na análise de importância dos atributos, conforme Figura 4, corrobora-se para os parâmetros na sessão de escolha das variáveis preditoras, e apresenta por ordem decrescente as variáveis mais importantes para o algoritmo, com destaque para representadas pelos indicadores “PL/Exigível Total”, “Lucros Retidos/Ativos”, “Liquidez Seca”, “Estoque/Ativos” e “Necessidade de Capital de Giro (NCG)”.

O *output* de um modelo não é gerado apenas pelas variáveis explicativas, mas também por meio da interação entre as variáveis. Sendo assim, realizou-se uma avaliação por

meio do cálculo do *H-statistic* de como o algoritmo utiliza a interação das variáveis para formar a previsão do modelo conforme proposto por Friedman e Popescu (2008).

A medida de interação diz respeito a quanto da variância de  $f(x)$  é explicada pela interação. A medida está entre 0 (sem interação) e 1 (= 100% da variância de  $f(x)$  devido às interações). Para cada variável, é medido o quanto ele interage com qualquer outra variável. O modelo realiza a interação das variáveis como forma de elevar sua capacidade preditiva.

Figura 5 - Interações das Variáveis

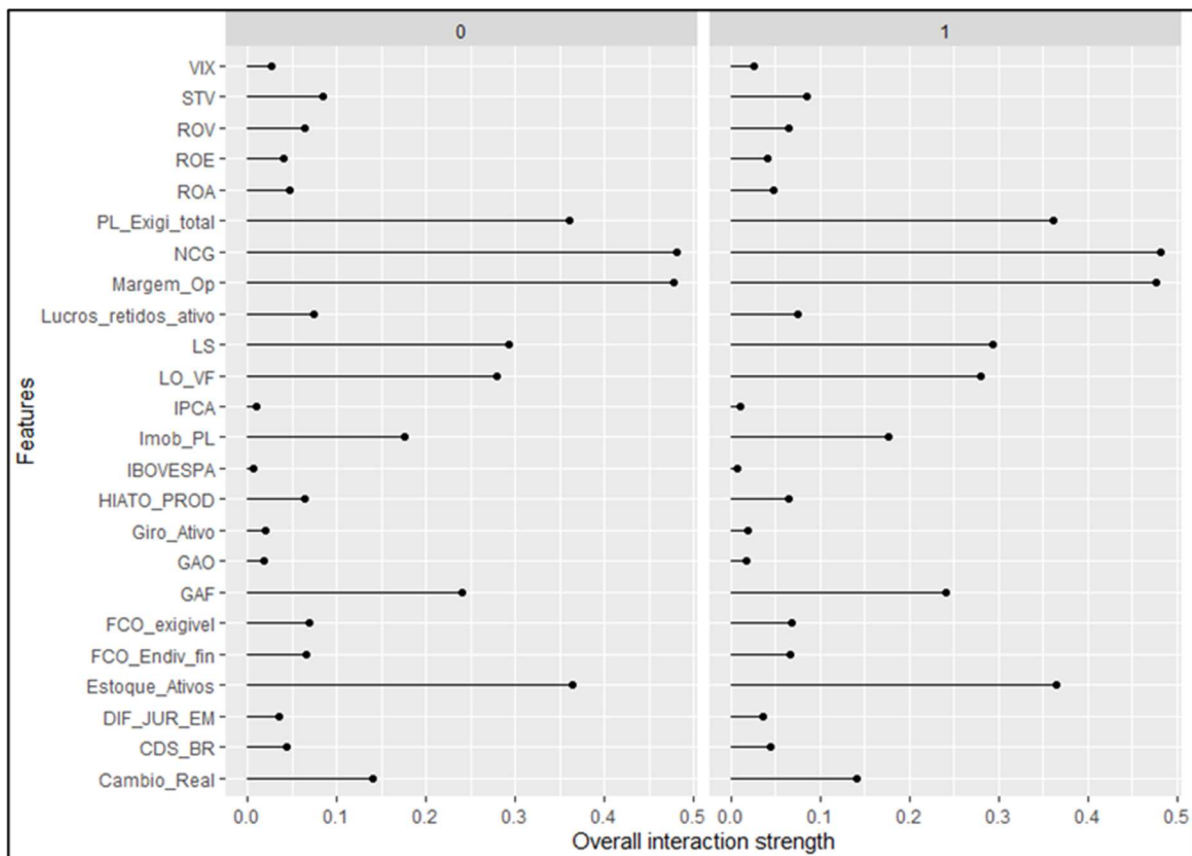


Figura apresenta a força de interação entre cada variável utilizada no modelo *XGBoost*, com as demais variáveis do modelo, para formulação da capacidade preditiva do algoritmo.

Fonte: Elaboração própria. Realizado em R – Pacote *lml*.

De maneira geral, pode se considerar que a interação entre as variáveis é importante na geração do *output* do modelo, em especial no que diz respeito às variáveis

“PL/Exigível Total”, “NCG”, “Lucros Retidos /Ativos” e “Estoque/Ativos”. Nesse caso, cabe um detalhamento de interação dessas variáveis, conforme apresentado a seguir. Vale destacar, ainda, que as variáveis de indicador zero explicam os valores previstos a partir de seus próprios valores, e não por meio de interações.

Deste modo, como se pode observar na Figuras 6, a variável PL/Exigível Total possui maior interação com as variáveis NCG, Lucros Retidos /Ativos e Liquidez Seca. No que se refere às interações que o modelo realiza com a variável NCG, esta possui maior interação com a variável Lucros Retidos/Ativos seguida pela PL/Exigível Total e Liquidez Seca e Estoque/Ativos, conforme apresentado na Figura 7.

**Figura 6 - Interação da variável PL/Exigível Total com as demais**

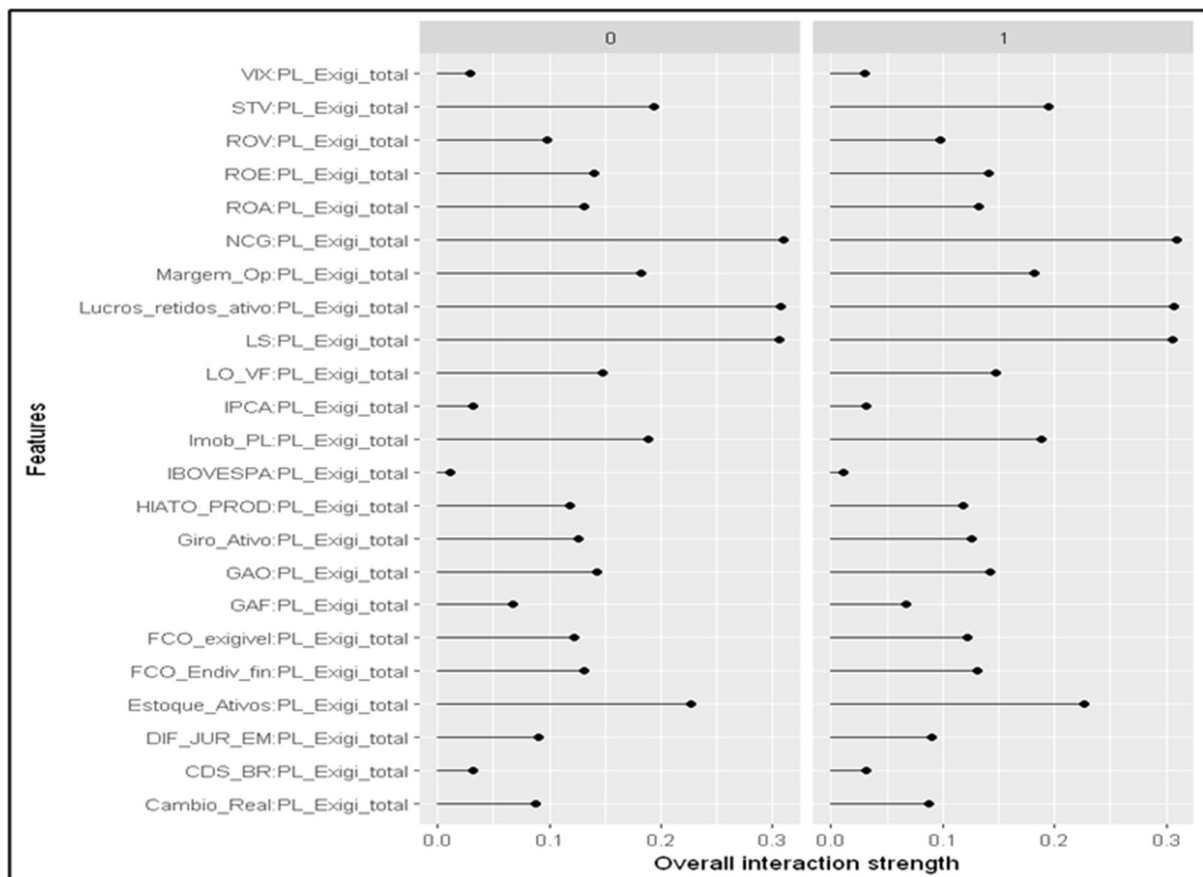


Figura apresenta a força de interação entre cada variável utilizada no modelo *XGBoost*, com a variáveis “PL/Exigível Total”, para composição da capacidade preditiva do algoritmo.

Fonte: Elaboração própria. Realizado em R – Pacote *Iml*.

Figura 7 - Interação da variável NCG com as demais

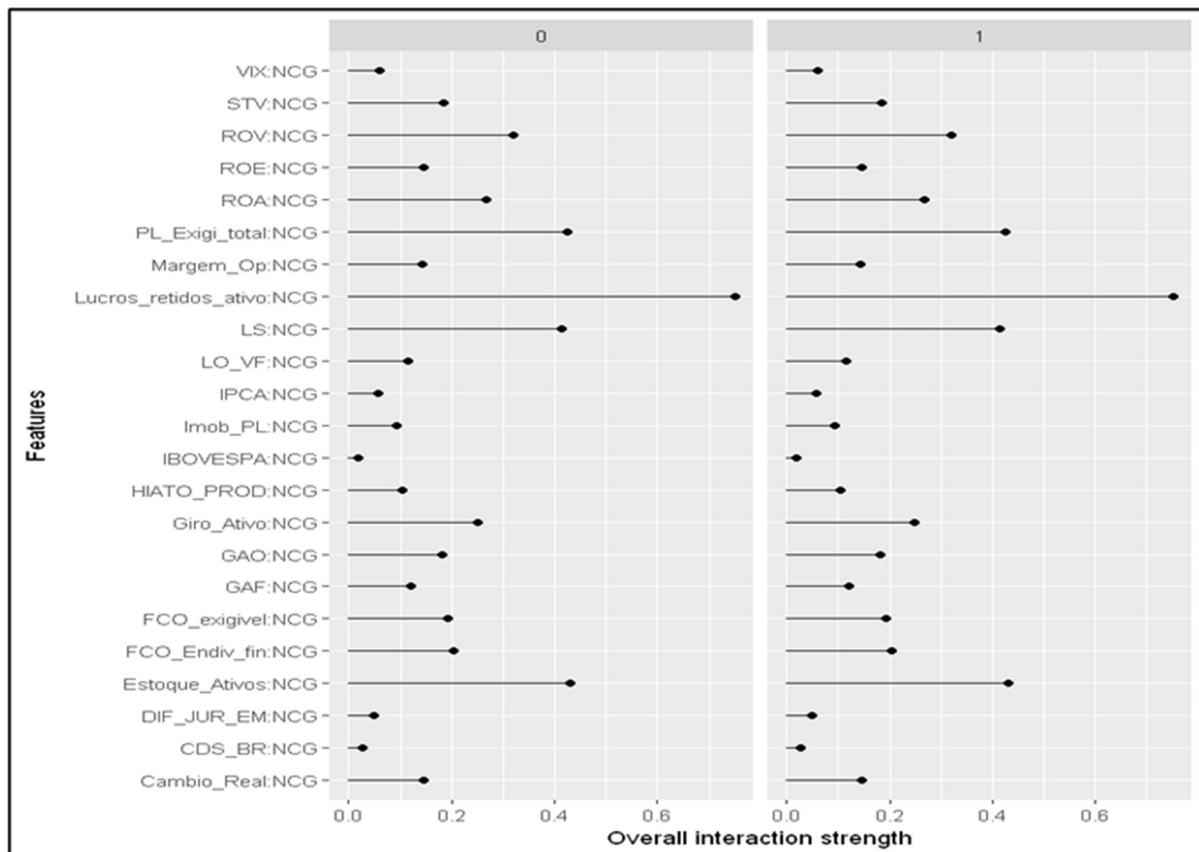


Figura apresenta a força de interação entre cada variável utilizada no modelo *XGBoost*, com a variáveis “NCG”, para composição da capacidade preditiva do algoritmo.

Fonte: Elaboração própria. Realizado em R – Pacote lml.

No que se refere a variável Margem Operacional, diferentemente das demais variáveis analisadas, a mesma realiza maior interação com a variável STV, seguido pelas variáveis Lucros Retidos sobre Ativos e Ibovespa conforme disposto na Figura 8.

Figura 8 - Interação da variável Margem operacional com as demais

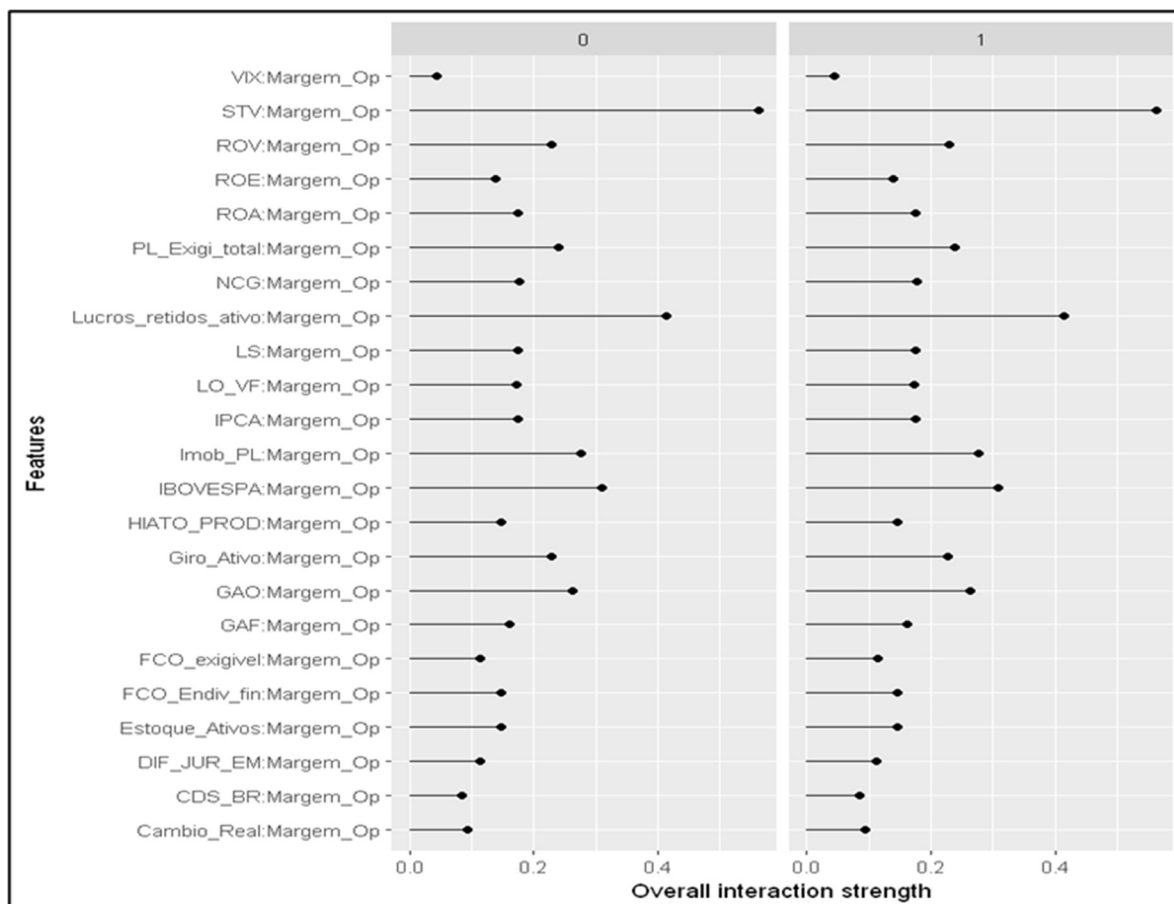


Figura apresenta a força de interação entre cada variável utilizada no modelo *XGBoost*, com a variáveis “Margem operacional”, para composição da capacidade preditiva do algoritmo.

Fonte: Elaboração própria. Realizado em R – Pacote lml.

Por fim, a variável Estoque/Ativos realizada maior interação com a Lucros Retidos/Ativos, Liquidez Seca e NCG conforme Figura 9.

Assim, verifica-se que as maiores interações destas quatro variáveis analisadas são entre elas mesmo em conjunto com a NCG. Deste modo, pode-se afirmar que essas são as variáveis mais importantes para a capacidade preditiva do modelo.



Figura 9 - Interação da variável Estoque/Ativos com as demais

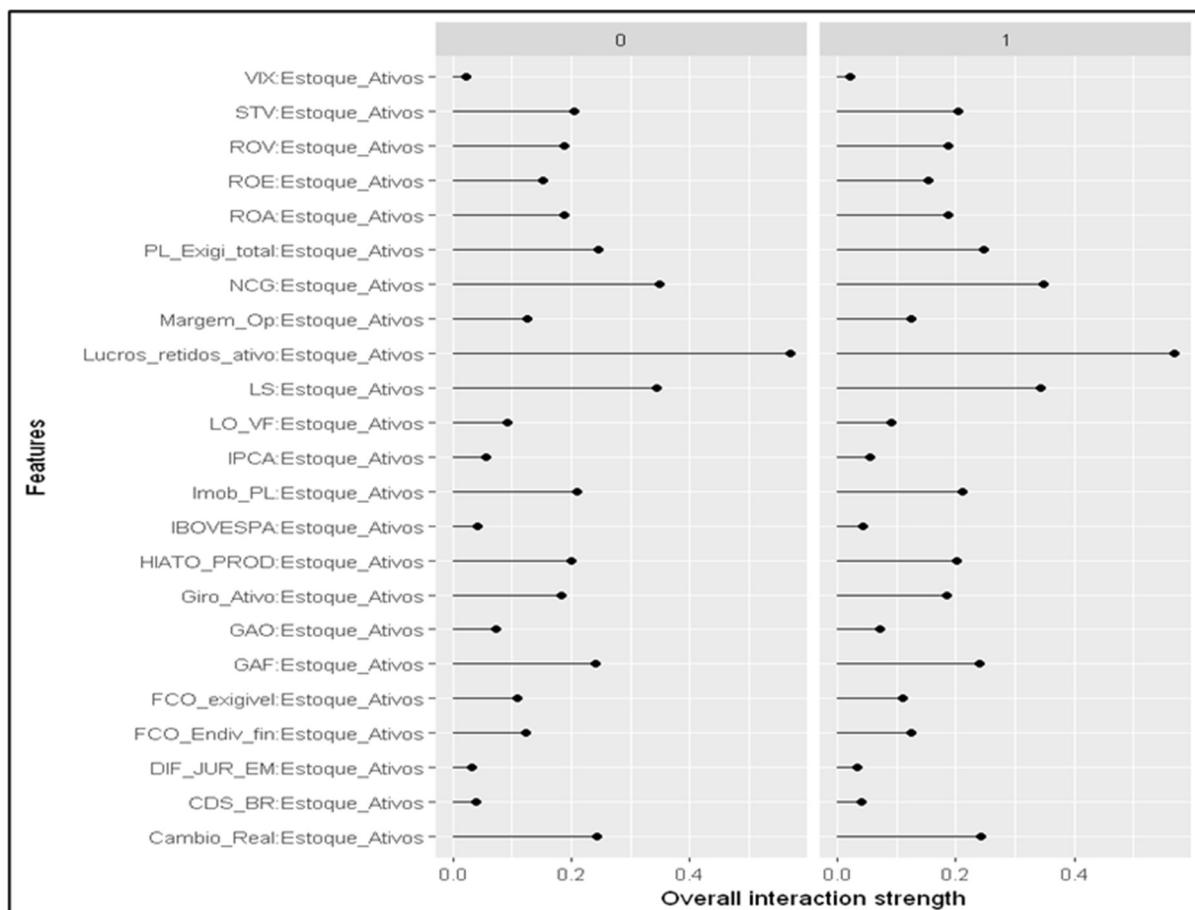


Figura apresenta a força de interação entre cada variável utilizada no modelo *XGBoost*, com a variáveis “Estoque/Ativos”, para composição da capacidade preditiva do algoritmo.

Fonte: Elaboração própria. Realizado em R – Pacote lml.

## 5. CONCLUSÃO

Por efeito da crescente disputa entre os bancos tradicionais e *fintechs* de todos os tipos e portes, cresce a importância para os agentes ofertantes de crédito, de que estes, cada vez mais, utilizem modelos robustos, que melhorem sua capacidade de estimação e avaliar o risco de crédito para diferentes perfis de clientes, tanto pessoas físicas quanto jurídicas.

O uso de técnicas de aprendizado de máquina, dado sua grande escala computacional e capacidade de reconhecimento de padrões, são cada vez mais importantes para uma melhor classificação dos clientes que a instituição financeira deseja trabalhar, dado sua capacidade de assunção de riscos.

Este trabalho buscou comparar diversos tipos de modelos para previsão do risco de crédito, mensurado pela probabilidade de descumprimento, para averiguar se os modelos com uso de técnicas de aprendizado de máquina, possuem melhor capacidade preditiva que os modelos ditos tradicionais. Para tanto, utilizou-se base de dados com informações macroeconômicas e dos balanços econômicos e financeiros de empresas do segmento atacado, disponibilizadas pela CVM, entre 2011 e 2019, somando-se um total de 7734 observações.

Deste modo, verificou-se que, a partir dos dados apresentados, a capacidade preditiva de alguns modelos de *machine learning* apresentam performance preditiva significativamente superior a tradicional regressão logística. Este resultado corrobora o verificado por Xia et al. (2017). Assim, o modelo *XGBoost* foi escolhido como aquele que apresentou a melhor performance (ROC de 0.99) dentre os modelos testados e também apresentou valor superior ao denotado por Brito e Assaf Neto (2009) com ROC de 0.97, ao qual obteve tal performance, contudo sem aplicar o modelo a uma amostra de observação fora da base de modelagem (*out-of-time*).

No que se tange as variáveis mais importantes para do modelo, as mais relevantes vão de encontro com o disposto na literatura ao apontar a escolha da empresa em sua formação de dívida, relação entre o patrimônio líquido e o passivo exigível total, a capacidade de reter lucros versus seus ativos totais e a gestão da liquidez da empresa, como o descrito por Soares e Rebouça (2015), Guimarães e Moreira (2008) e Jackson e Wood (2013).

No que se refere as avaliações de interpretabilidade do modelo, apresentou-se coerência entre a direção dos valores das variáveis e o sentido econômico das mesmas, bem como apontaram para a relação de variáveis mais importantes e suas principais interações, a

saber: PL/Exigível Total, Lucros Retidos/Ativos, Liquidez Seca, Estoque/Ativos e Necessidade de Capital de Giro (NCG). Verifica-se também que valor nominal da variável, sua grandeza, apresenta sentido econômico coerente para atribuição do valor justo (*shapley value*) em direção a classificação *default* ou *não-default*, corroborando, mais uma vez, para o sentido econômico do modelo.

Cabe destacar que, embora o *shapley value* e mais indicadores de interpretabilidade ainda sejam embrionárias e ainda pouco utilizadas nas finanças, bancos e supervisores para validação de modelos, fica evidente que tem-se expandido sua abrangência para interpretar algoritmos de “*black-box*” onde não são visíveis os pesos (betas) de cada variável na conjunção da predição final.

## REFERÊNCIAS

Banco Central Do Brasil – Bacen. (2009). **Resolução C.M.N. nº 3.721**. Dispõe sobre a implementação de estrutura de gerenciamento do risco de crédito.

Betancourt, G. A. (2005). **Las máquinas de soporte vectorial (SVMs)**. Scientia et technica, 1(27).

Breiman, L. (2001). **Random forests**. Machine learning, 45(1), 5-32.

Black, F., & Scholes, M. (1973). **The pricing of options and corporate liabilities**. Journal of political economy, 81(3), 637-654.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992, July). **A training algorithm for optimal margin classifiers**. In Proceedings of the fifth annual workshop on Computational learning theory (pp. 144-152).

Brito, G. A. S., Assaf Neto, A., & Corrar, L. J. (2009). **Sistema de classificação de risco de crédito: uma aplicação a companhias abertas no Brasil**. Revista contabilidade & finanças, 20(51), 28-43.

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, 321-357.

Chen, T., & Guestrin, C. (2016, August). **Xgboost: A scalable tree boosting system**. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Damodaran, A. (2010). **Avaliação de investimentos: ferramentas e técnicas para a determinação do valor de qualquer ativo**. Qualitymark.

Friedman, J. H., & Popescu, B. E. (2008). **Predictive learning via rule ensembles**. The Annals of Applied Statistics, 2(3), 916-954.

Guimarães, A., & Moreira, T. B. S. (2008). **Previsão de insolvência: um modelo baseado em índices contábeis com utilização da análise discriminante**. Revista de Economia Contemporânea, 12(1), 151-178.

Grmanová, E., & Ivanová, E. (2018). **Efficiency of banks in Slovakia: Measuring by DEA models**. Journal of International Studies Vol, 11(1).

Gu, S., Kelly, B., & Xiu, D. (2018). **Empirical asset pricing via machine learning** (No. w25398). National Bureau of Economic Research.

Haykin, S. (2007). **Redes neurais: princípios e prática**. Bookman Editora.

Hull, J. (2012). **Risk management and financial institutions**, Web Site (Vol. 733). John Wiley & Sons.

Jackson, R. H., & Wood, A. (2013). The performance of insolvency prediction and credit risk models in the UK: A comparative study. The British Accounting Review, 45(3), 183-202.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). **An introduction to statistical learning** (Vol. 112, p. 18). New York: springer.

Lewis, D. D. (1998, April). **Naive (Bayes) at forty: The independence assumption in information retrieval**. In European conference on machine learning (pp. 4-15). Springer, Berlin, Heidelberg.

Lundberg, S. M., & Lee, S. I. (2017). **A unified approach to interpreting model predictions**. In Advances in neural information processing systems (pp. 4765-4774).

Luo, C., Wu, D., & Wu, D. (2017). **A deep learning approach for credit scoring using credit default swaps**. Engineering Applications of Artificial Intelligence, 65, 465-470.

McCulloch, W. S., & Pitts, W. (1943). **A logical calculus of the ideas immanent in nervous activity**. The bulletin of mathematical biophysics, 5(4), 115-133.

Merton, R. C. (1974). **On the pricing of corporate debt: The risk structure of interest rates**. The Journal of finance, 29(2), 449-470.

Perlin, M. (2017). **GetDFPData: Reading Annual Financial Reports from Bovespa's DFP, FRE and FCA System. R package version 0.5**. Corporate disclosure quality and the cost of debt. Accounting review, 459-474, P.(1998). Disponível em: <https://CRAN.R-project.org/package=GetDFPDataSengupta>.

Pesaran, M. H., Schuermann, T., Treutler, B. J., & Weiner, S. M. (2006). **Macroeconomic dynamics and credit risk: a global perspective**. Journal of Money, Credit and Banking, 1211-1261.

Soares, R. A., & Rebouças, S. M. D. P. (2015). **Avaliação do desempenho de técnicas de classificação aplicadas à previsão de insolvência de empresas de capital aberto brasileiras**. Revista ADM. MADE, 18(3), 40-61.

Souza, Ê. B. M., & Corrar, L. J. (2010). **O Uso do Modelo de Merton para Obtenção de Spreads de Crédito: uma Proposta de Implementação Simplificada**. Sociedade, Contabilidade e Gestão, 5(1).

Tian, Y., Shi, Y., & Liu, X. (2012). **Recent advances on support vector machines research.** Technological and Economic Development of Economy, 18(1), 5-33.

Tukey, J. W. (1977). **Exploratory data analysis** (Vol. 2, pp. 131-160).

Xia, Y., Liu, C., Li, Y., & Liu, N. (2017). **A boosted decision tree approach using Bayesian hyperparameter optimization for credit scoring.** Expert Systems with Applications, 78, 225-241.

# conselho editorial

## COORDENAÇÃO

**José Luiz Rossi Junior**

**Thiago Costa Monteiro Caldeira**

## SUPERVISÃO E REVISÃO

**Renan Silveira Holtermann**

**Matheus Gonçalves**

**Mathias Schneid Tessmann**

**Milton de Souza Mendonça Sobrinho**

**Alessandro de Oliveira Gouveia Freire**

## COMUNICAÇÃO & MARKETING

**Antonio Aleixo Zaninetti Silva**

## PROJETO GRÁFICO E DIAGRAMAÇÃO

**Juliana Vasconcelos**

---

---



**idp**

SGAS Quadra 607 - Módulo 49  
Via L2 Sul, Brasília-DF  
CEP: 70200-670

  /sejaidp  
 (61) 3535-6565  
 idp.edu.br