INSTITUTO BRASILEIRO DE ENSINO, DESENVOLVIMENTO E PESQUISA

MESTRADO PROFISSIONAL EM ECONOMIA

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

Autor:Alex Cerqueira Pinto

Orientador: Prof. Dr. Alexandre Xavier Ywata de Carvalho

BRASÍLIA – DF

2020

ALEX CERQUEIRA PINTO

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

Dissertação apresentada ao programa do Mestrado Profissional em Economia Aplicada do Instituto Brasiliense de Direito Público, como parte dos requisitos necessários para a obtenção do Título de Mestre em Economia.

BRASÍLIA/ DF

2020

Dissertação de autoria de ALEX CERQUEIRA PINTO, intitulada “O poder preditivo dos modelos com aprendizado de máquina é superior aos modelos tradicionais para análise do risco de crédito? “, requisito necessário para a obtenção do grau de Mestre em Economia, defendida e aprovada em XX de novembro de 2020, pela banca examinadora constituída por:

_____

Prof. Dr. Alexandre Xavier Ywata de Carvalho
Orientador
Instituto Brasiliense de Direito Público

_____

Prof. Dr. Guilherme Mendes Resende
Examinador Interno
Instituto Brasiliense de Direito Público

_____

Prof. Dr. Daniel Oliveira Cajueiro
Examinador Externo
Universidade de Brasília

BRASÍLIA/ DF

2020

A tarefa não é tanto ver aquilo que ninguém viu, mas pensar o que ninguém ainda pensou sobre aquilo que todo mundo vê.” (Arthur Schopenhauer)

AGRADECIMENTOS

Em primeiro lugar, agradeço a Deus por permitir-me concretizar, nesse momento, mais um sonho de vida que consiste na busca contínua por aprendizado e conhecimento, objetivando crescer e me desenvolver pessoal e profissionalmente.

Em seguida, agradeço a minha família pelo amparo e incentivo em cada momento vivido. Estes que sempre me deram todo o suporte para poder voar e seguir o caminho que eu o desejar. Meu pai, com certeza, sei que torce por mim dia a dia e cada conquista minha sei que é dele também. À minha mãe, minha inspiração e meu exemplo, sempre esteve comigo em cada momento acadêmico e dedicando seu total apoio a minha trajetória, te dedico esse momento especial.

Agradeço a minha esposa e companheira incondicional Thais pela presença constante em todos os meus projetos, ao qual foi essencial para a conclusão deste trabalho. Grato pela sua compreensão com as minhas horas de ausência. Te amo. Esta é apenas uma das muitas conquistas ao seu lado. Somos um só.

Agradeço também a todo o corpo docente do Mestrado Profissional em Economia do IDP, em especial, ao Prof. Dr. Alexandre Ywata, exímio orientador, a quem tive a oportunidade de conhecer e conviver durante as aulas do curso e a jornada da dissertação. Igualmente, agradeço ao Prof. Mathias Tessamann, colaborador do IDP e monitor de tantas disciplinas cursadas ao longo destes dois anos de aprendizado, por sua capacidade de organização e comprometimento pela qualidade do curso.

Por fim, agradeço a colaboração de todos os colegas da turma do Mestrado em Economia com os quais tive o prazer e privilégio de conhecer e aprender, em especial aquela cuja amizade ultrapassou a fronteira do mestrado e se tornou uma grande amiga para vida, Raquel de Sá. Muito obrigado, também, ao meus companheiros e colegas do Banco do Brasil, que me apoiaram e deram suporte para que tudo isso acontecesse.

RESUMO

O objetivo desse trabalho foi desenvolver modelos para previsão do risco de crédito, para verificar se modelos com uso de aprendizado de máquina apresentam melhor caráter preditivo comparado a tradicional regressão logística. Do mesmo modo, como objetivo específico, aplicar técnicas de interpretabilidade ao modelo de melhor performance, A metodologia adotada corresponde a uma pesquisa empírica econométrica com o uso das técnicas de aprendizado supervisionado. O público alvo foram empresas do segmento atacado, que possuem registros na Comissão de Valores Mobiliários (CVM). Para as variáveis do modelo foram utilizados indicadores econômicos e financeiros, retirados das demonstrações contábeis e patrimoniais das empresas, e também variáveis macroeconômicas. Os resultados indicam que o modelo de melhor capacidade preditiva foi o XGBoost, com curva ROC na base teste de 0.99 e acurácia de 0.98 Do mesmo modo, as principais variáveis preditivas foram os indicadores de PL/Exigível Total, Lucros Retidos/Ativos, Liquidez Seca, Estoque/Ativos e Necessidade de Capital de Giro (NCG). Na análise de interpretabilidade via *Sharp value*, os resultados corroboram a intepretação da importância e sentido econômico das variáveis. Assim, o *Sharp value* indica uma relação inversa entre as variáveis PL/Exigível Total, Liquidez Seca e Lucros Retidos/Ativos e o valor predito. Do mesmo modo, a interpretabilidade via interações mostrou que, para o modelo, as variáveis PL/Exigível Total, Necessidade de Capital de Giro, Lucros Retidos /Ativos e Estoque/Ativos são as que apresentam interações mais fortes com as demais variáveis. Estes resultados corroboram a tendência de crescimento do uso dos modelos com uso de técnicas de *machine learning* na área econômica por, muitas vezes, apresentarem melhor capacidade preditiva.

Palavras-Chave: Risco de Crédito; Aprendizado de Máquina; Indicadores Financeiros; Modelo.

ABSTRACT

The purpose of this paper was to develop models for predicting credit risk, to verify if models using machine learning have a better predictive character compared to traditional logistic regression. Likewise, as a specific objective, to apply interpretability techniques to the best performance model. The methodology adopted corresponds to an empirical econometric research using supervised learning techniques. The target audience was wholesale companies, which have registrations with the Brazilian Securities and Exchange Commission (CVM). The model variables were economic and financial indicators, taken from the companies' accounting and equity statements, and macroeconomic information. The results indicate that the model with the best predictive capacity was the XGBoost, with a RCO curve at the test base of 0.99 and accuracy of 0.98 Likewise, the main predictive variables were the indicators of PL/Total Liabilities, Retained/Active Profits, Dry Liquidity, Inventory /Assets and Working Capital Need (NCG). In the analysis of interpretability via Sharp value, the results corroborate the interpretation of the importance and economic sense of the variables. Thus, the Sharp value indicates an inverse relationship between the variables Equity/Total Liabilities, Dry Liquidity and Retained Earnings/Assets and the predicted value. Likewise, the interpretability via interactions showed that, for the model, the variables PL/Total Liabilities, NCG, Retained Earnings/Assets and Inventory/Assets are those that present stronger interactions with the other variables, especially among themselves and together with the NCG and STV variables. These results corroborate the growth trend in the use of models with the use of machine learning techniques in the economic area, as they often have better predictive capacity.

Keywords: Credit Risk; Machine Learning; Financial Indicators; Model

SUMÁRIO

1. INTRODUÇÃO

A principal contribuição para a perenidade de uma Instituição Financiera 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)

Assim como em qualquer empresa, uma instituição financeira está sujeita a grande diversidade de riscos durante a condução de seus negócios. Conhecer suas características e particularidades é fundamental, já que os riscos aos quais está exposta, e que não sabe reconhecer, são os que se revelam mais contundentes (MARTIN**,** SANTOS, DIAS FILHO, 2004).

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. (ANTUNES et al., 2005)

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 et al. (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 BIS[1] nos chamados Acordos de Basiléia.

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 et al, 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 f*intechs*; ii) redução da taxa básica de juros no Brasil e no cenário global, que reduz os ganhos de tesouraria e pressionam o *spreed* 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. Este processo necessita de aprimoramento e desenvolvimento continuo, 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.

Para tanto, foi desenvolvido 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 curca ROC, para a escolha daquele com maior poder preditivo quanto a probabilidade de descumprimento (*default*).

[1] *Bank for International Settlements*, o Banco de Compensações Internacionais. Criado em 1930, o BIS é uma organização internacional que fomenta a cooperação entre os bancos centrais e outras agências em busca da estabilidade monetária e financeira, que propõe requisitos mínimos de capital para bancos conforme seus ativos ponderados pelo risco.

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[2] do modelo de melhor performance com o uso das técnicas a) *sharpe value*; b)Importâncias das variáveis; e c) Medidas das Interações *(Measure Interactions)*.

Este trabalho terá seu foco de atuação circuscrito 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.

Assim, o presente artigo está organizado em 4 sessões:

Sessão I - Introdução: Nesta primeira parte do projeto, foi feito a contextualização do tema de pesquisa, na sequência, o problema de pesquisa, em que se destaca a criação e mensuração de modelos de risco de crédito com o uso de técnicas estatísticas de aprendizado de máquina para previsão de descumprimento de empresas brasileiras de capital aberto. Este capítulo também apresenta os objetivos gerais e específicos do projeto, delimitação do escopo, justificativa da pesquisa e por último a organização do trabalho.

Sessão II – Fundamentação Teórica: nesta parte do projeto é apresentado o modelo teórico mais usualmente utilizado para risco de crédito e mensuração de probabilidade de descumprimento. Também é apresentada a revisão da literatura nacional e internacional que subsidia a aplicação empírica sobre o tema.

Sessão III – Metodologia: Neste capítulo é apresentado a metodologia de pesquisa, os contextos metodológicos e os métodos científicos. Também é apresentado a forma de coleta, descrição e fonte dos dados, técnica quantitativa (econométrica) a ser utilizada, bem como a delimitação da população e amostra do objeto de estudo.

Sessão IV –Resultados: neste capítulo é descrito os resultados do processo de modelagem, os indicadores de cada modelo e os resultados da predição.

[2] Para realizar intepretabilidade, utilizou-se o pacote SHAP em linguagem Python e o pacote IML - Interpretable Machine Learning na linguagem R.

2. FUNDAMENTAÇÃO TEÓRICA

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 contra posição 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:

$$E_0 = A_0.N(d_1) - D.e^{-rT}.N(d_2)$$

,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 juro 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:

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

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 e ASSAF NETO, 2009)

Matematicamente a regressão logística é descrita como:

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

Onde α = é constante da equação e β = são os coeficientes das variáveis preditoras.

2.2.Literatura Nacional e Internacional

A literatura nacional e internacional 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, os mesmos 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 et al. (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 et al. (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 Autores Revisão Bibliográfica.

Autores	Técnica Utilizada	Público alvo	Dados	Conclusões
Brito e Assaf Neto (2009)	Regressão Logística	60 empresas de capital aberto no	Índices financeiros calculados a partir das demonstrações	Curva ROC de 0,978 e quatro variáveis explicativas, a saber: i)

		período entre 1994 e 2004.	contábeis do penúltimo exercício anterior ao ano default.	lucros retidos sobre ativo; ii) endividamento financeiro; iii) capital de giro líquido; e iv) saldo de tesouraria sobre vendas.
Soares e Rebouça (2015)	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 default 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.
Guimarães e Moreira (2008)	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.
Jackson e Wood (2013)	Total de 13 modelos de *machine learning* 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.
Luo et al. (2017)	*Deep Learnig*, Regressão Logística Multinomial (MLR), Perceptron Multilayer (MLP) e SVM.	Rating de crédito de empresas globais, que atuam em diversos setores, após a crise de 2008.	Usaram como variável o *Credit Defaul Swap* (CDS) das empresas.	Os resultados indicam que o uso de *Deep Learning* supera outros algoritmos e podem facilmente classificar o risco de crédito antecipadamente das empresas.
Xia et al. (2017)	XGBoost	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 superam 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 da *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 aprendizado supervisionado é fornecido um conjunto de dados rotulados onde é informado a saída/categorização correta do objeto. Assim apresenta-se a ideia de que existe uma relação entre as características de entradas e a saídas. No modelo de risco, para as características extraídas de cada variável da amostra indicou-se ao modelo se a empresa era “boa” ou “ruim”. Em síntese, modelos supervisionados apresentam a variável dependente (*target*) nos dados de treinamento do modelo. O objetivo do modelo é reconhecer um padrão de comportamento para classificação. (GREGÓRIO, 2018)

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 sobreajustamente (*overfiting*) 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	(Ativo Circulante + Realizável a Longo Prazo) / (Passivo Circulante + Exigível a Longo Prazo)
V2	Liquidez corrente	Ativo Circulante / Passivo Circulante
V3	Liquidez seca	(Ativo Circulante – Estoques) / Passivo Circulante
V4	Liquidez imediata	Disponível / Passivo Circulante
V5	Retorno sobre o patrimônio líquido	Lucro Líquido / Patrimônio Líquido inicial
V6	Retorno sobre o ativo	LAJIR / Ativo Total
V7	Retorno sobre vendas	Lucro Líquido / Vendas Líquidas
V8	Giro do ativo	Vendas Líquidas / Ativo Total
V9	Margem operacional	LAJIR / Vendas Líquidas

V10	Lucro operacional sobre despesas financeiras	LAJIR / DF
V11	Patrimônio líquido sobre ativo	Patrimônio Líquido / Ativo Total
V12	Lucros retidos sobre ativo	(Lucros Acumulados + Reserva de Lucros) / Ativo Total
V13	Patrimônio líquido sobre exigível total	Patrimônio Líquido / (Passivo Circulante + Exigível a longo Prazo)
V14	Endividamento total	(Passivo Circulante + Exigível a longo Prazo) / Ativo Total
V15	Endividamento de curto prazo	Passivo Circulante / Ativo Total
V16	Endividamento financeiro	(Passivo Circulante Financeiro + Exigível a longo Prazo Financeiro) / Ativo Total
V17	Imobilização do patrimônio líquido	Ativo Permanente / Patrimônio Líquido
V18	Estoques sobre ativo	Estoques / Ativo Total
V19	Capital de giro líquido	(Ativo Circulante – Passivo Circulante) / Ativo Total
V20	Necessidade de capital de giro	(Ativo Circulante Operacional – Passivo Circulante Operacional) / Ativo Total
V21	Saldo de tesouraria sobre ativo	(Ativo Circulante Financeiro – Passivo Circulante Financeiro) / Ativo Total
V22	Saldo de tesouraria sobre vendas	(Ativo Circulante Financeiro – Passivo Circulante Financeiro) / Vendas Líquidas
V23	Fluxo de caixa operacional sobre ativo	Fluxo de Caixa das Operações / Ativo Total
V24	Fluxo de caixa operacional sobre exigível total	Fluxo de Caixa das Operações / (Passivo Circulante + Exigível a longo Prazo)
V25	Fluxo de caixa operacional sobre endividamento financeiro	Fluxo de Caixa das Operações / (Passivo Circulante Financeiro + Exigível a longo Prazo Financeiro)

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

Tabela 3 - Relação de Variáveis Macroeconômicas

Código	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 Cambio - Real/Dólar - Nominal	Numérica
M16	Taxa de Cambio - 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, 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 et. al. (2002) e aplicado na linguagem R.

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 et al. (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 dado 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:

$$P(C = c_k/X = x) = \ P(C = c_k)x\frac{P(X = x/C = c_k)}{P_{(x)}}$$

$$P(X = x) = \sum_{k'=1}^{ec} P(X = x/C = c_{k'})\, x\, P(C = c_{k'})$$

, 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 o 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 “o” (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:

- Construção das bases de treinamento utilizando *bootstrap* na base de treinamento original. Amostragem com reposição para formação dos dados;
- Criar múltiplos algoritmos construídos para cada conjunto de dado reamostrado;
- Combinar os algoritmos: As predições são combinadas utilizando médias, moda, mediana para regressão ou voto majoritário para problemas de classificação.

No caso da *Random Forest*, este modelo combina várias arvores 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 et al, 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 e 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.

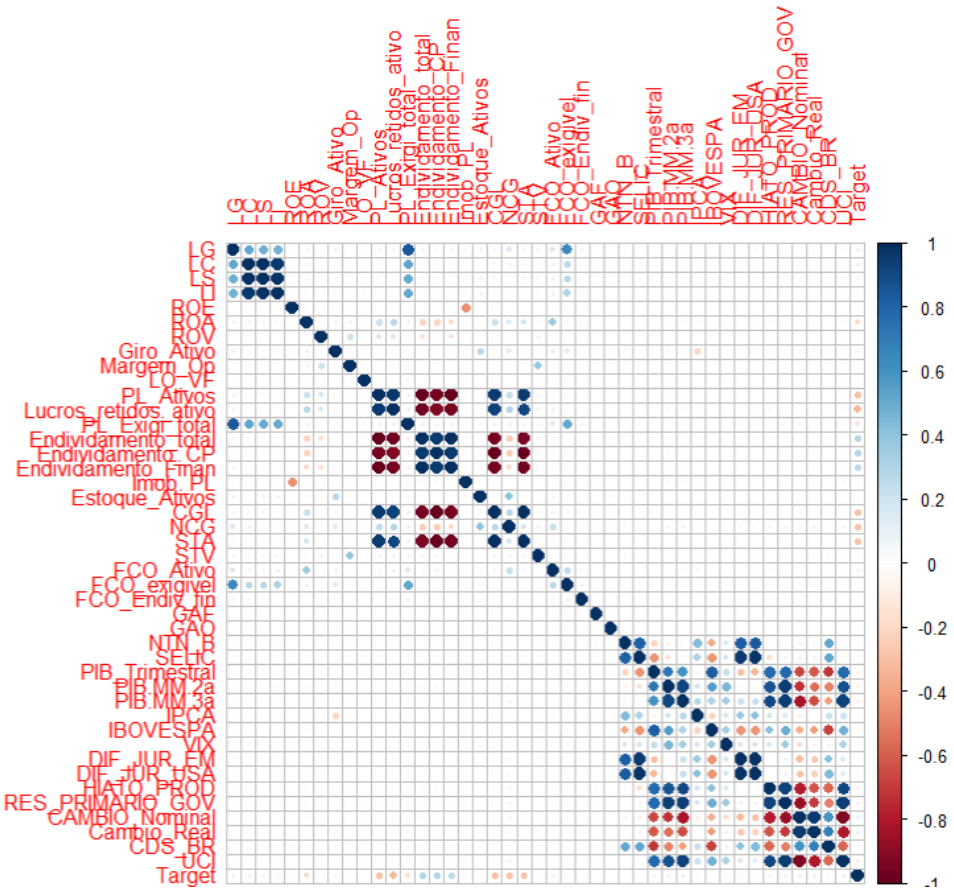
4. RESULTADOS

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.

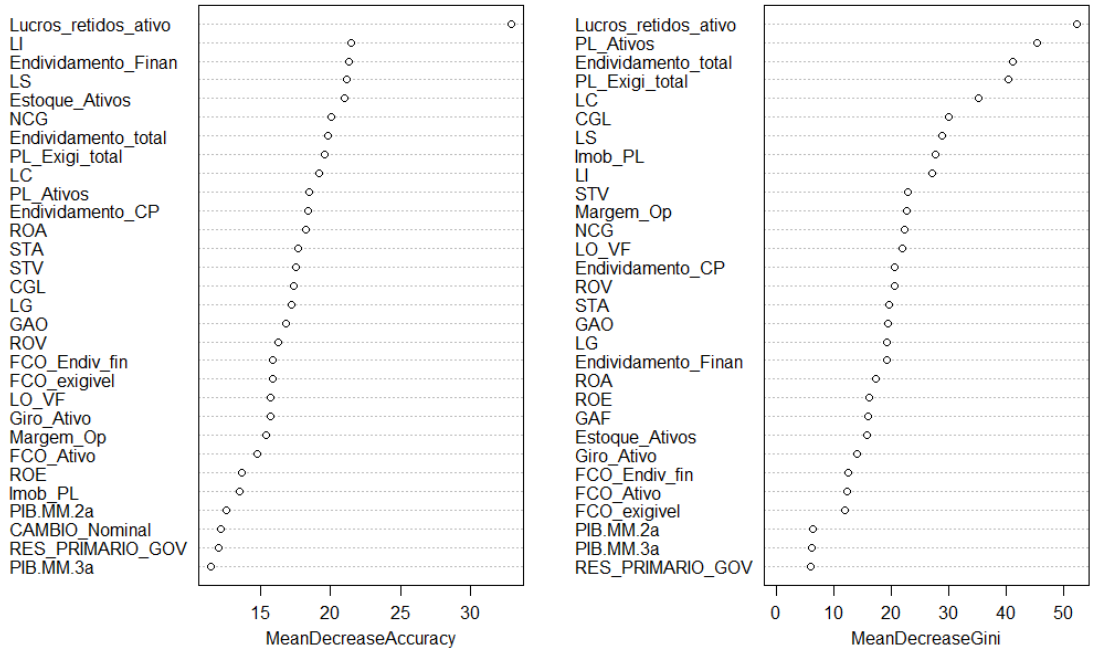
De forma adicional para escolha das variáveis, foi realizada avaliação via importância relativa das variáveis, através da técnica de árvore de decisão, para selecionar aquelas com maior poder preditivo aos modelos conforme destaca a figura 2.

Figura 1 – Matriz de Correlação Pearson das variáveis preditoras.



Fonte: Base de Dados Modelagem. Elaboração própria

Figura 2 – Importância Relativa das Variáveis.



Fonte: Base de Dados Modelagem. 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 3 abaixo:

Tabela 3 – Relação de Variáveis Finais para uso nos modelos.

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

Fonte: Base de Dados Modelagem. Elaboração própria

4.2.Comparativo dos Modelos e escolha da melhor performance

Após a aplicação aos dados as técnicas mencionadas, utilizando todos os modelos com a melhor configuração de hiperparametros possível [3], verificacou-se que os algoritimos 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 0.97com os dados balanceados.

[3] A busca pelos melhores hiperparametros dos modelos foi realizada por intermédio da função *Grid Search* em R. Mais informações disponíveis em: https://medium.com/fintechexplained/what-is-grid-search-c01fe886ef0a

Tabela 5 – Indicadores de performance dos algoritimos – Dados em nível.

	Naive Bayes	SVM - Linear	SVM - Polynomial	SVM - Radial	Regressão Logística	Redes Neurais	Random Forest	XgBoost
Acurácia	0,9496	0,9528	0,9509	0,9651	0,9655	0,9638	0,978	0,9838
Precision	0,6666	0,6538	0,9	0,8478	0,851	0,7258	0,9615	1
Recall	0,097	0,2098	0,1071	0,4534	0,3539	0,5357	0,6097	0,66
F1	0,17	0,3177	0,1914	0,5909	0,5	0,6164	0,7462	0,7967
ROC	0,89	0,91	0,94	0,94	0,94	0,96	0,97	0,99

Fonte: Elaboração própria

Tabela 6 – Indicadores de performance dos algoritimos – Dados Balanceados.

	Naive Bayes	SVM - Radial	Redes Neurais	Regressão Logística	SVM - Polynomial	SVM - Linear	XgBoost	Random Forest
Acurácia	0,9347	0,9244	0,9166	0,9024	0,9257	0,9011	0,9586	0,9606
Precision	0,4054	0,3629	0,355	0,3251	0,3856	0,3236	0,5625	0,5798
Recall	0,5625	0,6125	0,75	0,825	0,7375	0,8375	0,90	0,8625
F1	0,4712	0,4558	0,4819	0,4664	0,5064	0,4669	0,6923	0,6934
ROC	0,87	0,88	0,90	0,91	0,91	0,92	0,97	0,97

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 intepretabilidade de suas estimativas. Modelos com esse tipo de técnica são considerados do tipo "*black-box* por se tratar de um algoritimo não tradicional onde não se pode observar os valores dos betas estimados para as variáveis, bem como realizar testes estatísticos padrões.

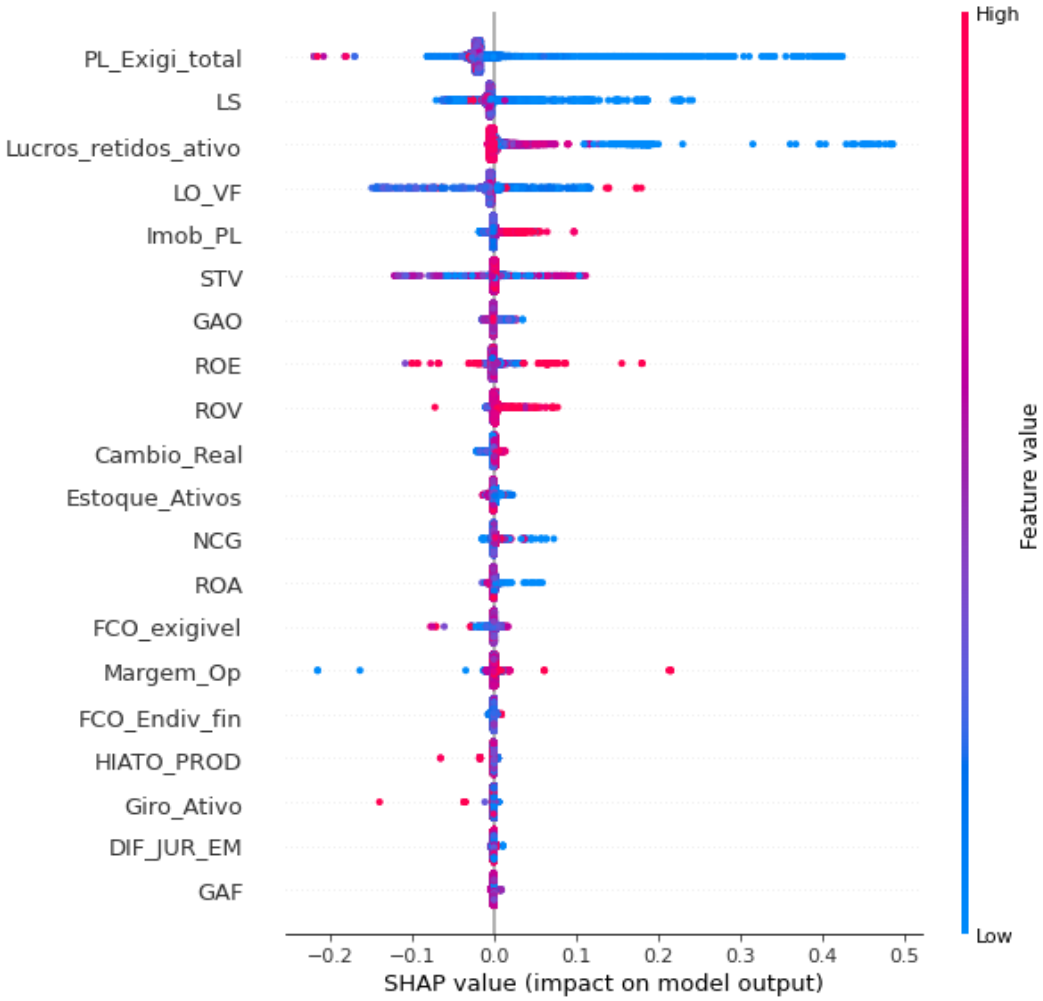
Os modelos de machine learning, conhecidos como "*black-box*", são cada mais estudados com o objetivo de clarificar a relação entre as variáveis explicativas e o *output* do modelo. Importante ressaltar que a interpretabilidade do modelo visa, inclusive, avaliar a qualidade do modelo, à medida que permite 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) *Sharpley Value*; b) Importância e c) Interactions.

A partir do conceito de *Sharpley Value*, oriundo da Teoria dos Jogos, Lundberg et al. (2017) desenvolveram um método aplicado aos modelos de machine learning que confere interpretabilidade aos modelos. Neste contexto, o *sharpley value* mede a contribuição de cada variável na construção do *output*, ou seja, o valor justo que cada

variável influência 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 – Shap Value.

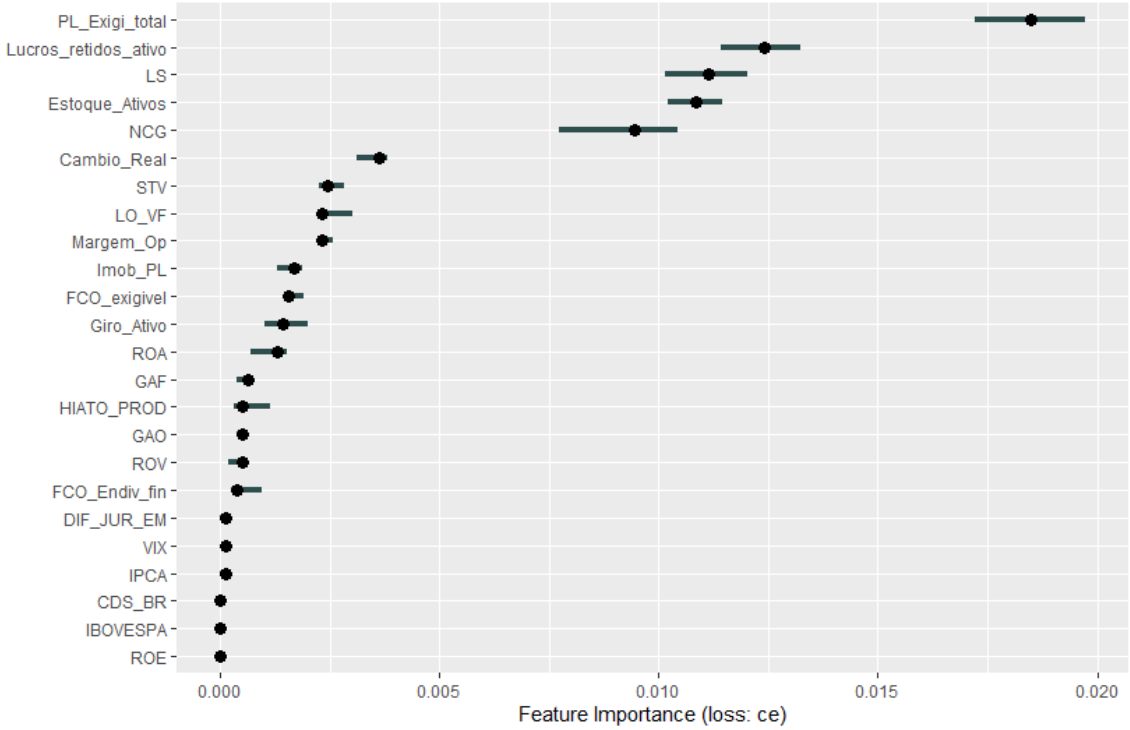


Fonte: Elaboração própria. Realizado em Python – Pacote Shap.

O gráfico da Figura 3 ordena as variáveis por importância, e corrobora os resultados apresentados pelo modelo no que se refere a intepretação econômica dos resultados. Deste modo, o *Sharpley 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 em azul) 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.



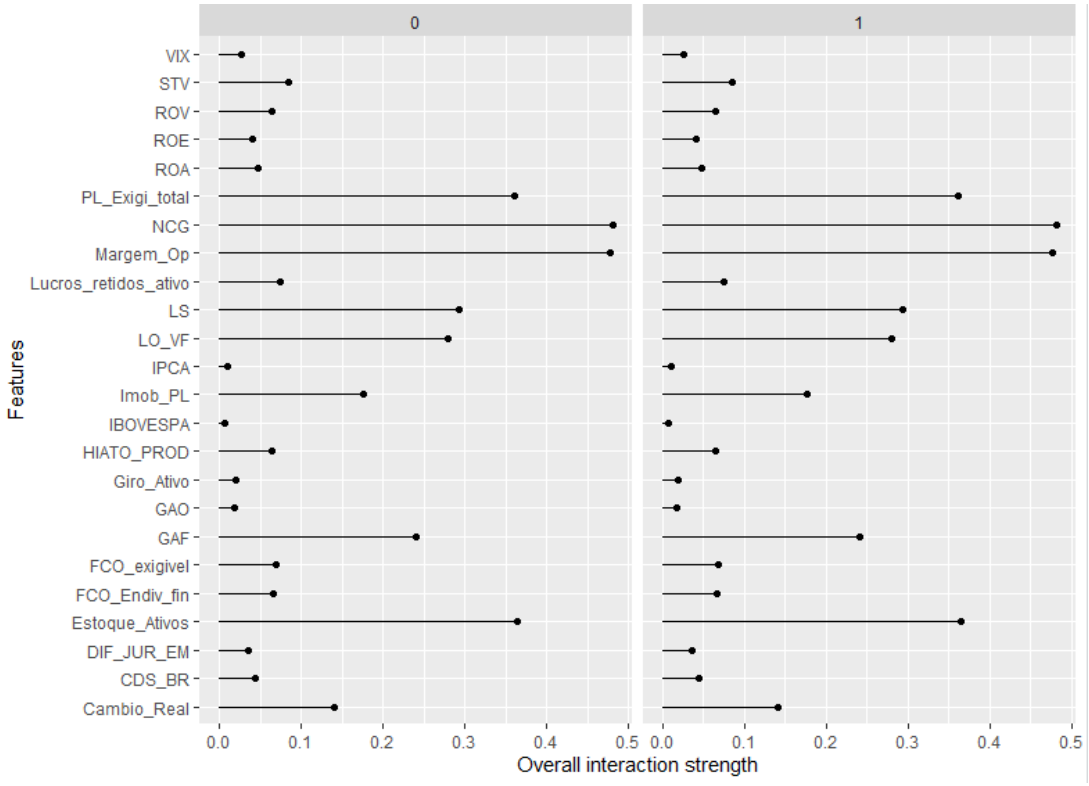
Fonte: Elaboração própria. Medida de erro: *Classification Error.* Realizado em R – Pacote Iml.

Na análise de importância dos atributos, conforme Figura 4, a mesma corrobora para os parâmetros na sessão de escolha das variáveis preditoras, e apresenta por ordem decrescente as variáveis mais importantes para o algoritimo, 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 algorítimo 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 outro variável. O modelo realiza a interação das variáveis como forma de elevar sua capacidade preditiva.

Figura 5 – Interações das Variáveis.



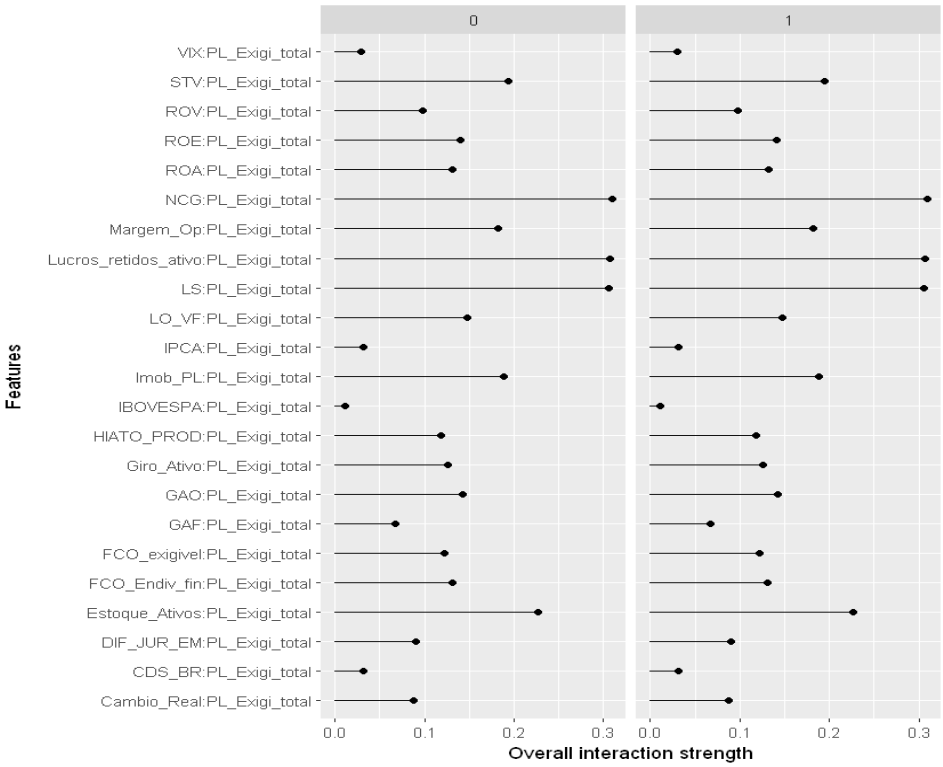
Fonte: Elaboração própria. Realizado em R – Pacote Iml.

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.

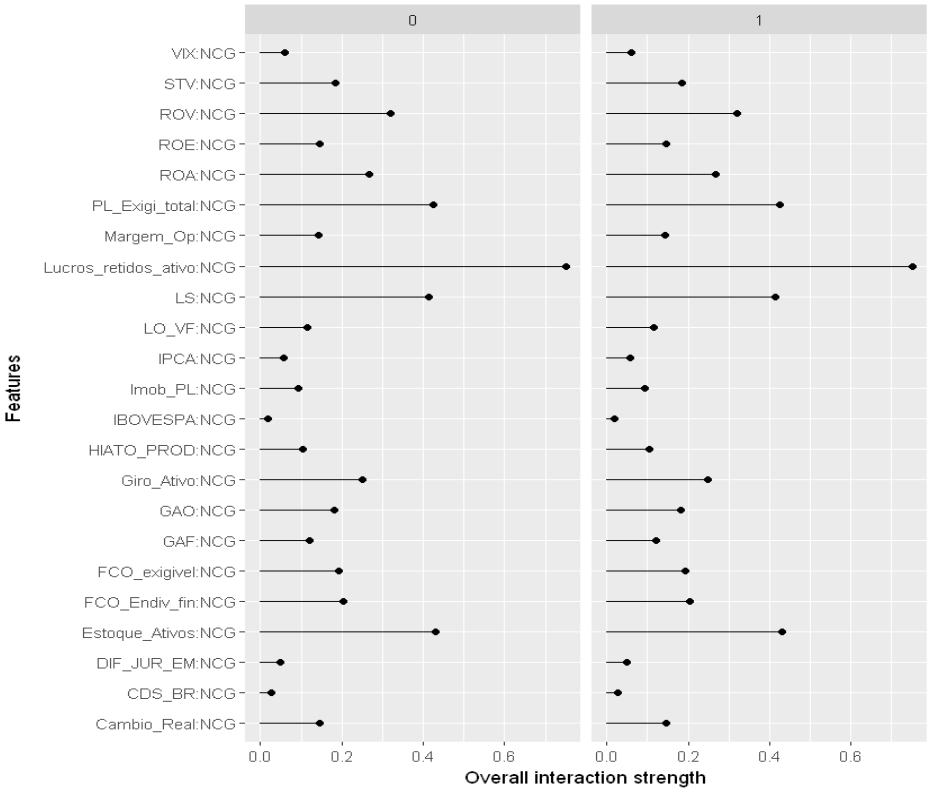
No que se refere as interações que o modelo realiza com a variável NCG, a mesma possui maior interação com a variável Lucros Retidos/Ativos seguida pela PL/Exigível Total e Liquidez Seca e Estoque/Ativos, conforme apresentado na Figura 7.

Figura 6 – Interação – PL/Exigível Total



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

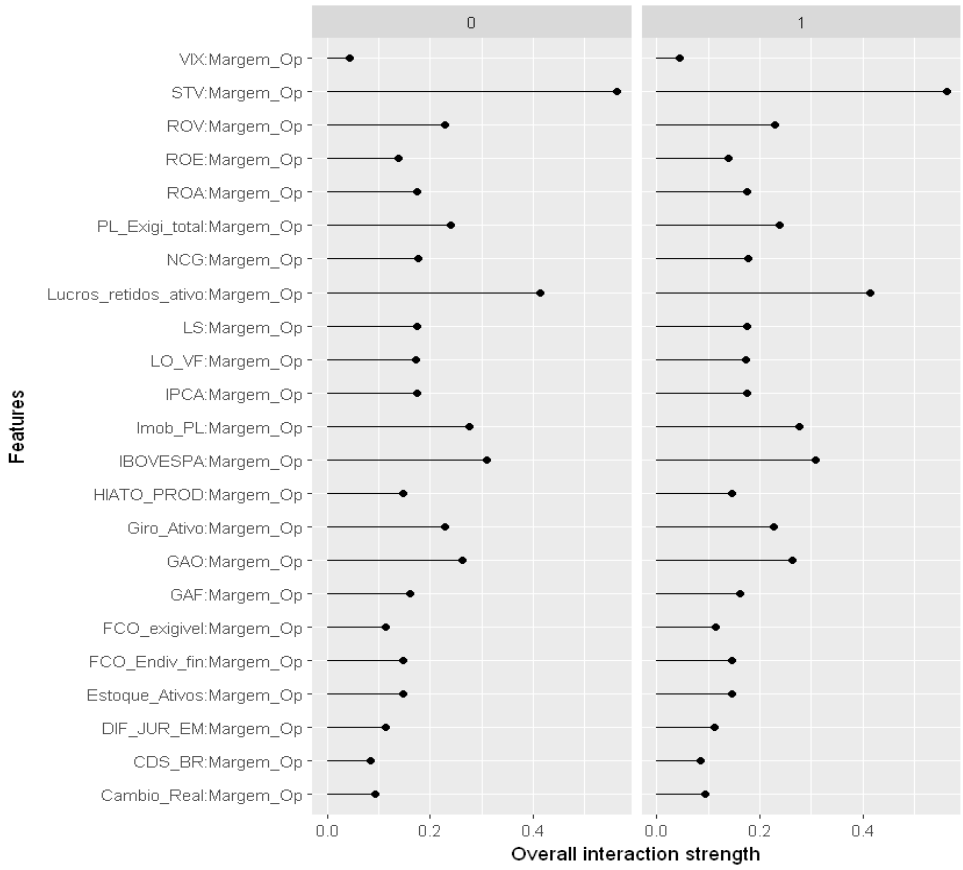
Figura 7 – Interação - NCG



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

No que se refere a variável Margem Operacional, diferentemente das demais variáves 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 – Margem operacional

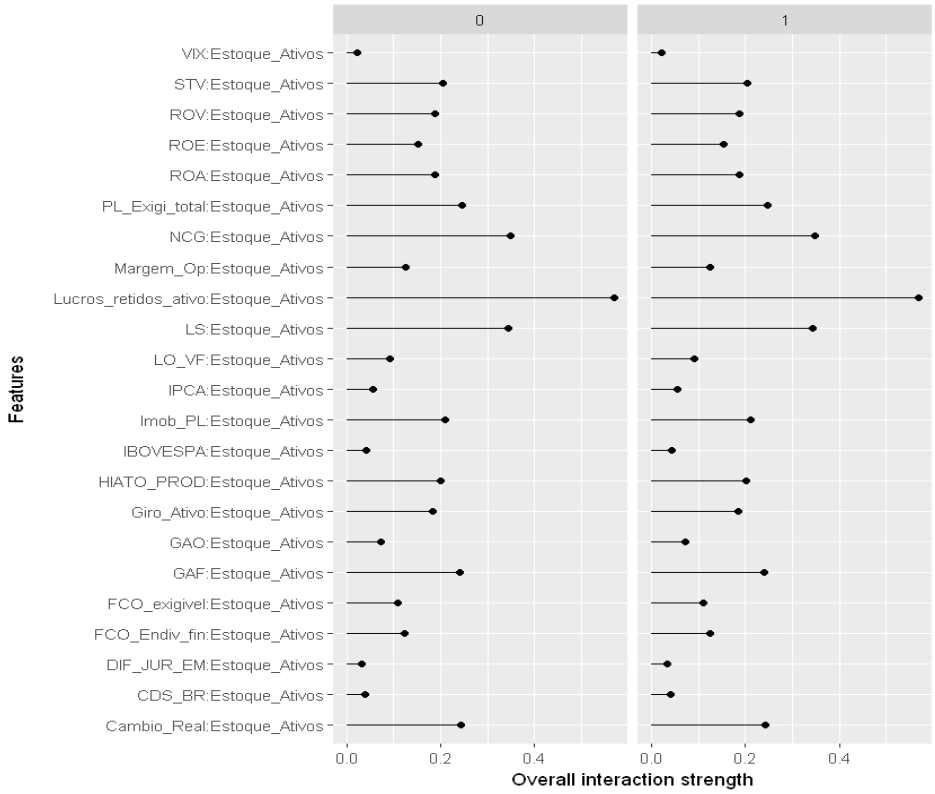


Fonte: Elaboração própria. Realizado em R – Pacote Iml.

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 as mesmas são as variáveis mais importantes para a capacidade preditiva do modelo.

Figura 9 – Interação - Estoque/Ativos



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

5. CONCLUSÃO

Por efeito da crescente disputa entre os bancos tradicionais e *fintechs* de todos os tipos e portes, cresce a impotância para os agentes ofertantes de crédito, de que estes, cada vez mais, utilizem modelos robustos, que melhorem sua capacidate 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. 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.

No que se refere as avaliações de interpretabilidade do modelo, os mesmos apresentaram 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).

Cabe destacar que, embora o *sharpley 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 as mesmas têm expandido sua abrangência para interpretar algoritmos de "*black-box*" onde não são visíveis os pesos (betas) de cada variável na conjunção da predição final.

6. REFERÊNCIAS

BANCO CENTRAL DO BRASIL – BACEN. **Resolução nº 3.721.** Dispõe sobre a implementação de estrutura de gerenciamento do risco de crédito, 2009.

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

BREIMAN, L. Random forests. **Machine learning**, v. 45, n. 1, p. 5-32, 2001.

BLACK, F., SCHOLES, M. The pricing of options and corporate liabilities. **Journal of political economy**, v. 81, n. 3, p. 637-654, 1973.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: **Proceedings of the fifth annual workshop on Computational learning theory**. p. 144-152, 1992.

BRITO, G. A. S.; ASSAF NETO, A.; CORRAR, L. J. Sistema de classificação de risco de crédito: uma aplicação a companhias abertas no Brasil. **Rev. contab. financ**. São Paulo, v. 20, n. 51, p. 28-43, Dec. 2009.

CHAWLA, N. V. et al. SMOTE: synthetic minority over-sampling technique. **Journal of artificial intelligence research**, v. 16, p. 321-357, 2002.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining**. p. 785-794, 2016.

DAMODARAN, A. **Avaliação de investimentos: ferramentas e técnicas para a determinação do valor de qualquer ativo**. 3ª Reimpressão. Qualitymark, 2010.

FRIEDMAN, J H; POPESCU, B. E. Predictive learning via rule ensembles. **The Annals of Applied Statistics,** v. 2, n. 3, p. 916-954, 2008.

GUIMARÃES, A.; MOREIRA, T. B. S. Previsão de insolvência: um modelo baseado em índices contábeis com utilização da análise discriminante. **Revista de Economia Contemporânea**, v. 12, n. 1, p. 151-178, 2008.

GU, S.; KELLY, B.; XIU, D. **Empirical asset pricing via machine learning.** National Bureau of Economic Research, 2018.

HAYKIN, S. **Redes neurais: princípios e prática**. Bookman Editora, 2007.

JACKSON, R. H. G.; WOOD, A. The performance of insolvency prediction and credit risk models in the UK: a comparative study. **The British Accounting Review**, v. 45, p. 183-202, 2013.

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

LUNDBERG, Scott M.; LEE, Su-In. **A unified approach to interpreting model predictions**. In: Advances in neural information processing systems. 2017. p. 4765-4774.

LUO, C.; WU, D.; WU, D.. A deep learning approach for credit scoring using credit default swaps. **Engineering Applications of Artificial Intelligence**, v. 65, p. 465-470, 2017.

MERTON, R.C.; On The Pricing of Corporate Debt: The Risk Structure of Interest Rates. **The Journal of Finance**, 29, p. 449-470, 1974.

PERLIN, M. GetDFPData: Reading Annual Financial Reports from Bovespa's Dfp and Fre System. 2017. Disponível em https://github.com/msperlin/GetDFPData/. Acessado

em 13/03/2020.

PESARAN, M. H. et al. Macroeconomic dynamics and credit risk: a global perspective. **Journal of Money, Credit and Banking**, v. 38, n. 5, p. 1211-1261, 2006.

SOARES, R. A.; REBOUÇAS, S. M. D. P.. Avaliação do desempenho de técnicas de classificação aplicadas à previsão de insolvência de empresas de capital aberto brasileiras. **Revista ADM. MADE**, v. 18, n. 3, p. 40-61, 2015.

SOUZA, E. B. M.; CORRAR, L. J. O Uso do Modelo de Merton para Obtenção de Spreads de Crédito: uma Proposta de Implementação Simplificada. **Sociedade, Contabilidade e Gestão**, v. 5, n. 1, 2010.

TIAN, Y.; SHI, Y.; LIU, X. Recent advances on support vector machines research. **Technological and Economic Development of Economy**, v. 18, n. 1, p. 5-33, 2012.

TUKEY, J. W. **Exploratory data analysis**. 1977.

XIA, Y. et al. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. **Expert Systems with Applications**, v. 78, p. 225-241, 2017.