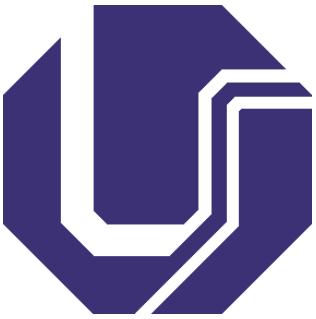


UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE GESTÃO E NEGÓCIOS
PROGRAMA DE PÓS-GRADUAÇÃO EM ADMINISTRAÇÃO
GESTÃO FINANCEIRA E CONTROLADORIA

PREVISÃO DE DIFICULDADES FINANCEIRAS EM EMPRESAS LATINO-AMERICANAS VIA APRENDIZAGEM DE MÁQUINA

VINICIUS NOGUEIRA MARRA

Orientador: Prof. Dr. Flávio L. M. Barboza

UBERLÂNDIA

2019

VINICIUS NOGUEIRA MARRA

PREVISÃO DE DIFICULDADES FINANCEIRAS EM EMPRESAS LATINO-AMERICANAS VIA APRENDIZAGEM DE MÁQUINA

Dissertação apresentada ao Programa de Pós-Graduação em Administração da Faculdade de Gestão e Negócios da Universidade Federal de Uberlândia, como requisito parcial para obtenção do título de Mestre em Administração.

Área de Concentração: Gestão Organizacional

Linha de Pesquisa: Gestão Financeira e Controladoria

Orientador: Prof. Dr. Flávio L. M. Barboza

UBERLÂNDIA

2019

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da UFU, MG, Brasil.

M358p
2019

Marra, Vinicius Nogueira, 1983-
Previsão de dificuldades financeiras em empresas latino-americanas via aprendizagem de máquina [recurso eletrônico] / Vinicius Nogueira Marra. - 2019.

Orientador: Flavio Luiz de Moraes Barboza.
Dissertação (mestrado) - Universidade Federal de Uberlândia, Programa de Pós-Graduação em Administração.
Modo de acesso: Internet.
Disponível em: http://dx.doi.org/10.14393/ufu.di.2019.947
Inclui bibliografia.
Inclui ilustrações.

1. Administração. 2. Administração de risco. 3. Aprendizado por computador. 4. Administração de empresas - Aspectos econômicos. I. Barboza, Flavio Luiz de Moraes, 1980-, (Orient.) II. Universidade Federal de Uberlândia. Programa de Pós-Graduação em Administração. III. Título.

CDU: 658

Gloria Aparecida - CRB-6/2047

Reitor da Universidade Federal de Uberlândia
Valder Steffen Júnior

Diretora da Faculdade de Gestão e Negócios
Kárem Cristina de Sousa Ribeiro

Coordenadora do Programa de Pós-Graduação
Cíntia Rodrigues de Oliveira Medeiros

VINICIUS NOGUEIRA MARRA

PREVISÃO DE DIFICULDADES FINANCEIRAS EM EMPRESAS LATINO-AMERICANAS VIA APRENDIZAGEM DE MÁQUINA

Uberlândia, 28 de fevereiro de 2019.

Banca Examinadora:

Prof. Dr. Flávio L. M. Barboza
Universidade Federal de Uberlândia

Prof. Dr. Antônio Sérgio Torres Penedo
Universidade Federal de Uberlândia

Prof. Dr. Pedro Henrique Melo Albuquerque
Universidade de Brasília

Com amor aos meus pais e minha esposa.

Aos meus pais, incomensurável gratidão; à minha esposa, todo carinho e respeito.

AGRADECIMENTOS:

A presente dissertação (de mestrado) não poderia chegar ao ideal de um bom porto sem o precioso apoio de várias pessoas as quais agregaram conhecimento e sabedoria a minha construção enquanto ser humano. Com elas aprendi que **devemos** ser gratos, (principalmente) sobretudo, por não nos dar tudo que lhe pedimos. Recordo-me com respeito e serenidade de todas as dificuldades as quais enfrentei; não fosse por elas, eu não teria (saído do lugar) transformado a inércia em força direcionada. Mesmo as críticas do Prof. Valdir que, logo no começo, me auxiliaram muito a expandir meus limites.

Em primeiro lugar, não posso deixar de agradecer ao meu orientador, Professor Doutor Flávio Luiz de Moraes Barboza, por toda a paciência didática, empenho meticuloso e rigidez prática com que sempre me orientou neste trabalho do mestrado.

Desejo igualmente agradecer a todos os meus colegas (do Mestrado) de curso - Nadjara, Valter, Debora, Kleverson, Dermeval, Juliana, João e aos Guilhermes – pela parceria e companheirismo. De modo especial, agradeço à Debora, que se (tornou) fez (grande) inestimável parceira na elaboração dos trabalhos das disciplinas. Agradeço a todos os técnicos administrativos do PPGA, especialmente a Juliana, pela atenção e disponibilidade dispendida às necessidades burocráticas.

Por último, quero agradecer à minha família e amigos pelo apoio incondicional que me deram, (especialmente) sobretudo a minha esposa por toda paciência genuína para comigo ao longo da elaboração deste trabalho.

“A teoria sem a prática vira 'verbalismo', assim como a prática sem teoria, vira ativismo. No entanto, quando se une a prática com a teoria tem-se a práxis, a ação criadora e modificadora da realidade.”

Paulo Freire

RESUMO

A capacidade de prever dificuldades financeiras nos negócios é primordial, pois decisões quanto a inapropriadas concessões de crédito podem ter consequências financeiras diretas e indiretas em toda economia. A previsão de falências e a mensuração do escore de crédito são dois importantes tópicos de pesquisa, tanto no campo contábil quanto no financeiro e este estudo almejou apresentar a direção das pesquisas acadêmica sobre a gestão de risco de crédito sob a ótica da aprendizagem de máquina. O interesse pelo estudo simultâneo desses dois tópicos surgiu após notar que ainda há resquícios e sentimentos da última crise financeira, mesmo uma década após seu acontecimento. Aliado a esse fato, o crescente desenvolvimento e aplicação da Inteligência Artificial às finanças prometem aumentar o rigor e aprimoramento de processos de análise de informações financeiras independentemente de quem seja o cliente. Por meio de um processo de seleção de artigos cientificamente relevantes no campo de pesquisa escolhido, obteve-se uma amostra final de 168 estudos considerados hábeis a nortear os rumos do uso da aprendizagem de máquinas – uma subárea da Inteligência Artificial - aplicados às finanças. A partir dessa amostra de estudos, foi realizada uma análise mais profunda por meio de suas leituras e descobriu-se que os algoritmos de aprendizagem de máquina estão sendo explorados, aprimorados e levados ao extremo para detectar combinações sutis e melhor descrever o risco de crédito. A partir dos resultados e lacunas encontrados na pesquisa, neste estudo foi proposto o uso de um modelo de algoritmo baseado em aprendizagem de máquinas para buscar a melhor previsão de acordo com os dados previamente fornecidos. Xgboost é um modelo que surgiu em meados de 2016 e vem ganhando notoriedade devido a sua acurácia e sua optimização computacional. Por ser um modelo relativamente novo, estudos com sua aplicação em finanças, especificamente, não foram encontrados. Neste contexto, o modelo proposto foi capaz de apresentar melhores resultados, quando comparados com Regressão Logística e *Random Forest*. Os resultados encontrados foram interessantes, pois identificaram outras lacunas e possibilidades de pesquisas futuras. A intenção de comprovar que o uso de inteligência artificial agrega valor à análise de credito é demonstrado.

Palavras-chave: *Extreme Gradient Boosting*, análise de risco de crédito, aprendizagem de máquina.

ABSTRACT

The ability to foresee financial distress in business is paramount, as decisions regarding inappropriate credit concessions may have direct and indirect financial consequences throughout the economy. Predicting bankruptcies and measuring credit scores are two important research topics, in both accounting and finance, and this study aims to demonstrate the direction of academic researches on credit risk management from the perspective of machine learning. Interest in simultaneously examining these two topics emerged after noting that remnants and feelings of the last financial crisis still lingers in society, even a decade after it occurred. Paired with this fact, the increasing development and application of Artificial Intelligence in finance promises to increase the rigor and improvement of financial information analysis processes. Through a structured process to select scientifically relevant articles in the chosen research field, a final sample of 168 studies deemed able to guide the directions of use of machine learning - sub-area of Artificial Intelligence - applied to finance was obtained. From this sample of studies, a deeper analysis was carried out through their readings and it was identified that machine learning algorithms are being explored, improved and taken to extreme to detect subtle combinations and better describe credit risk. From the results and gaps found in the research, in this study proposed the use of a novel algorithm used to teach machines to seek the best classification according to the data previously provided. Xgboost is a model that appeared in 2016 and has been gaining notoriety due to its accuracy and its computational optimization. Because it is a relatively new model, studies with its application in finance, per se, have not been found. Thus, in this context, the proposed model was able to present better results when compared to Logistic Regression and Random Forest. The results were interesting as they identified other gaps and possibilities for future research. The intention to prove that the use of artificial intelligence adds value to credit analysis has been demonstrated and, in this way, it is believed that by advancing this research, interesting results should be achieved, both for suppliers, analysts, managers, and investors.

Keywords*: Extreme Gradient Boosting, credit risk analysis, machine learning.*

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

Sumário

CAPÍTULO 1: INTRODUÇÃO

1.1 Contextualização do tema e problema de pesquisa

Desde a crise financeira global em 2008, o crescimento econômico mundial começou a desacelerar. Em um ambiente econômico instável, empresas sofrem grande pressão, e muitas delas caem em dificuldades financeiras ou mesmo em falências. Consequentemente, a dificuldade financeira das empresas tende a aumentar o risco do setor bancário por serem os maiores fornecedores de crédito do mercado (SUN *et al*, 2017). Nessa dinâmica há uma necessidade de quantificar e traduzir as informações do cliente na forma de risco envolvido em suas operações.

A melhora contínua desse processo de avaliação é de extrema importância, pois o Acordo de Basileia II exige que essas instituições financeiras estejam os mais seguros e respaldados possíveis para que possam continuar concedendo credito. A última crise de crédito – Crise Imobiliária Americana de 2008 - que o mundo vivenciou fez com que autoridades questionassem a questão da regulamentação envolvidos na emissão de crédito e os perigos envolvidos. (KRUGMAN, 2008)

Hoje essa análise ainda envolve a habilidade de pessoas na tomada de decisão final de crédito, dentro de um cenário de incertezas e constantes mutações e informações incompletas (SCHRICKEL; 2000). Ou seja, parte da análise é realizada por meio de modelos estatísticos que se apoiam integralmente na hipótese de ergodicidade e, com isso, pode-se identificar um certo padrão de comportamento (GALINDO e TAMAYO, 2000). O julgamento do agente de crédito, baseada principalmente na habilidade e experiência do mesmo ainda é o fator determinante na concessão do crédito final.

Todavia, modelos tradicionais carecem da habilidade de lidar, de maneira eficaz, com a não-linearidade dos dados. A partir desta vulnerabilidade, pesquisadores tem buscado modelos dinâmicos que possam se adequar à não-linearidade e aprender com essas particularidades dos dados para que o mínimo de informação seja perdido e a capacidade de mensurar o risco seja mais eficaz. (FAYYAD et al., 1996; BIGUS, 1996; ADRIAANS; ZANTINGE, 1996).

A Aprendizagem de Máquina ou *Machine Learning* é o ramo da ciência da computação que lida com o desenvolvimento de programas de computador que ensinam e se desenvolvem. De acordo com Arthur Samuel, o programador "dá ao computador a capacidade de aprender sem ser programado explicitamente". (SAMUEL, 1959). Dessa interpretação, pode-se deduzir que a Aprendizagem de Máquina permite que os desenvolvedores criem algoritmos que se

aprimoram automaticamente encontrando padrões nos dados existentes sem instruções explícitas de um humano ou desenvolvedor.

Em vista disso, esse trabalho não tem a intenção de questionar essas práticas de análise ou apontar falhas, mas sim, investigar se a aplicação de métodos computacionais de análise de dados - que automatizam a construção de modelos assintóticos - podem agregar valor a essa tomada de decisão, visto que um julgamento possível de identificar se uma empresa possui idoneidade ou capacidade financeira suficiente para amortizar a dívida que se pretende contrair pode ser tendenciosa e abstrata.

1.2 Objetivos do estudo

Tendo em vista a importância da gestão de risco para a economia e seus *stakeholders*, este trabalho utilizou-se de informações financeiras de empresas latino-americanas para realizar o estudo. O objetivo geral consistiu em analisar se a Aprendizagem de Máquina, representada pelo modelo proposto, agrega valor à análise e confrontá-los com modelos usuais nesse processo de gestão de risco.

De forma específica, os objetivos do estudo são: (1) Analisar se a adoção de modelos computacionais agregam à gestão de risco corporativo; (2) Examinar se a Aprendizagem de Máquina oferece uma maior precisão na previsão de dificuldades financeiras de empresas e (3) Verificar o quão precisos são modelos econométricos tradicionais quando comparados a modelos computacionais.

1.3 Importância do tema e justificativa do estudo

A previsão de dificuldades financeiras continua sendo uma importante área de enfoque para os pesquisadores, empresas e partes interessadas, incluindo investidores, credores e participantes do mercado de capitais em geral (WAQAS; MD-RUS, 2018). Investimentos financeiros, tais como ações, futuros, opções, são os mais conhecidos na sociedade. Contudo, mercados de capitais são voláteis e a maioria dos investidores sabe que a empresa está com problemas financeiros somente após a declaração pública do evento (CHEN; DU, 2009). Portanto, a previsão de dificuldades financeira de empresas desempenha um papel cada vez mais importante na sociedade, dado seu impacto nas decisões de empréstimo e a rentabilidade das instituições financeiras.

Esse desafio de prever tais eventos em corporações é importante pois, contratempos financeiros tornam-se onerosos uma vez que cria uma tendência de empresas a tomarem decisões que são prejudiciais para *debtholders* e *stakeholders* (partes interessadas não-

financeiras, tais como: clientes, fornecedores e funcionários), prejudicando o acesso ao crédito e elevando os custos de relações com partes interessadas (OPLER; TITMAN, 1994).

Pesquisas anteriores mostram que, quando algumas empresas do mesmo setor têm dificuldades financeiras, os custos do financiamento externo aos concorrentes aumentam (BENMELECH; BERGMAN, 2011; HERTZEL; OFFICER, 2012). Como consequência, custos de financiamento mais elevados em um setor poderiam reduzir o investimento, afetando a capacidade dos concorrentes de obter fundos necessários (OPLER; TITMAN, 1994).

Nesse contexto, instituições financeiras buscam aprimorar seus modelos e formas de analises de concessão de crédito e o Acordo Basiléia II visa melhorar a solidez desse complexo sistema. Esse acordo instituiu diretrizes regulatórias que dão mais ênfase aos controles internos dos próprios bancos para o gerenciamento de riscos. Principalmente após a Crise de 2008, onde uma reação em cadeia, gerou prejuízos em todo o mundo, o reconhecimento aos variados níveis de complexidade e sofisticação dos bancos internacionais, fez com que o Acordo Basiléia II ficasse mais em destaque. Sua resolução permite que bancos ofereçam uma estrutura flexível de seus próprios sistemas para medir seus riscos de mercado e, em última instância, gerenciar seus negócios com mais eficiência. (BIS, 2006).

A Basileia II oferece uma gama de metodologias para a medição de risco de crédito e risco operacional para que os bancos possam adotar abordagens que melhor se ajustem ao seu perfil de risco. Ao mesmo tempo, o acordo exige, também, uma divulgação abrangente por parte dos bancos de seus procedimentos de análise de risco, os quais estão sujeitos a revisão e avaliação por parte dos responsáveis pela supervisão do acordo.

É nesse contexto que este estudo busca apresentar um modelo de gestão de risco que ensaia unir algumas dessas exigências, as quais são: (1) o modelo proposto é de domínio público, logo, está aberto a investigações e eventuais avaliações; (2) a revisão literária do assunto em questão indica que o algoritmo traz consigo as tendências que buscam melhorar a acurácia da predição e oferece oportunidades para futuros estudos; (3) sua construção maximiza o uso do sistema computacional, evitando, assim, a demora no tempo de resposta e otimiza o (na otimização do) processo; (4) o modelo em questão indica quais variáveis tem um maior peso na predição da resposta e a partir desse último aspecto, os bancos podem desenvolver um sistema de alerta para prever quando seus clientes estão se aproximando de um limiar de risco.

1.4 Estrutura do trabalho

Esta dissertação está dividida em quatro capítulos. Após este capítulo introdutório, são desenvolvidos dois artigos independentes nos dois próximos capítulos. O capítulo 2 trata da revisão literária sistêmica da gestão de risco de crédito sob a ótica do Aprendizado de Máquina, enquanto o capítulo 3 discorre sobre algumas lacunas encontradas na revisão pela perspectiva (sob a ótica) de modelos de aprendizagem de máquina. O quarto capítulo apresenta as conclusões da pesquisa em questão.

CAPÍTULO 2: APLICAÇÃO DE APRENDIZAGEM DE MAQUINAS NA GESTÃO DE RISCO - UMA REVISÃO TEMÁTICA DA LITERATURA

Quando um número significante de devedores se torna incapaz de pagar empréstimos devido a dificuldades financeiras, o risco de crédito pode se espalhar por todo mercado, caracterizando uma cadeia de eventos negativos em todo o sistema financeiro, um efeito dominó na econômica. A Crise Financeira de 2008, como exemplar de um evento sistêmico dessa magnitude, ocasionou ondas de falências e desemprego em todo o mundo, e foi desencadeada por empréstimos de alto risco feitos no mercado imobiliário americano (KRUGMAN, 2008). O Fundo Monetário Internacional (FMI) estimou que bancos e outras instituições financeiras sofreram perdas financeiras de cerca de 4,1 trilhões de dólares como resultado da crise (FMI, 2009).

Na década de 80, mercados já se mostravam cautelosos com a concessão de crédito e o sistema bancário. Assinaram então, o acordo de Basileia que apresenta princípios básicos para uma metodologia de avaliação de risco de crédito que busca conciliar liquidez e estabilidade financeira aos que seguiam suas regras. Em 2005 reforçaram as diretrizes no Acordo de Basileia II, que visava fornecer incentivos aos bancos para adotarem sistemas quantitativos de gestão de risco baseados em dados, que poderiam promover a estabilidade sistêmica e financeira (BIS, 2006).

Em um artigo publicado pela McKinsey and Company (2015), a empresa de consultoria afirmou que a gestão de risco passará por mudanças substanciais nas próximas décadas, devido às regulamentações emitidas após a crise financeira global. Conforme relatado no estudo, um dos principais avanços que transformarão a análise de risco será derivada de técnicas de inteligência artificial.

Embora haja uma imensidade de métodos estatísticos e de inteligência artificial disponíveis, ainda não há um consenso quanto a melhor estratégia de análise financeira (DESAI et al., 1996; LEONARD, 1996; THOMAS, 1998; WEST, 2000; BAESENS, 2004, HUANG et al., 2004; LEE; CHEN, 2005; LENSBERG et al., 2006; BANASIK; CROOK, 2007; PALIWAL; KUMAR, 2009; CHEN et al., 2017; DIRICK et al., 2017).

Assim, este capítulo busca avaliar como a literatura sobre risco de crédito sob a ótica da aprendizagem de máquina tem (evoluiu) evoluído nos últimos 10 anos desde a Crise Financeira Global de 2008; assim como (e) sintetizar seu atual estado da arte e em qual direção está seguindo.

Para tal fim, utilizou-se o método ProKnow-C (*Constructivist Knowledge Process*) para seleção de estudos, que produziu uma seleção de 168 publicações científicas

internacionais. A partir desta investigação, pode-se confirmar que a aplicação da inteligência artificial (IA), neste campo, vem ganhando atenção especial. Notou-se, que a necessidade de identificar e incluir diferentes variáveis, melhorar a interpretabilidade dos resultados e aplicar empiricamente os modelos desenvolvidos são tendências e lacunas nesta área de estudo.

2.1 Literatura e abordagem geral

A segunda premissa do Acordo de Basileia II estabelece que os processos de revisão e supervisão dos métodos de análise de risco, é garantir que os bancos tenham capital adequado para resistir a todos os riscos relevantes a seus negócios e encorajar as instituições financeiras a adotarem técnicas sofisticadas de monitoramento e gestão de risco (BIS, 2006).

É importante ressaltar que a modelagem de risco de crédito tem sido foco de pesquisa há décadas. Altman (1968) usou Análise Discriminante Múltipla para classificar empresas estado de falência e sem tal risco e concluiu que a falência poderia ser explicada de forma bastante eficaz por meio de uma combinação de cinco índices financeiros (ex: capital de giro / ativos totais, lucros retidos / ativos totais, lucro antes de juros e impostos / total de ativos, valor de mercado do patrimônio / total de passivos e vendas / ativos totais). Ohlson (1980) aplicou modelos de Regressão Logística para prever dificuldades financeiras, enquanto Frydman, Altman e Kao. (1985) utilizaram árvores de decisão.

Mais recentemente, técnicas estatísticas avançadas, juntamente com métodos de inteligência artificial, foram integradas com o objetivo de lidar com essa não-linearidade. Por exemplo, *Support Vector Machine* (HUANG et al., 2004; VAN GESTEL et al. 2006; LI et al., 2012), Redes Neurais Artificiais (TSENG; HU, 2010; ALP et al., 2011; KHASHMAN, 2011; BLANCO et al., 2013) e Técnicas de Pesquisas Evolucionárias (ONG; HUANG; TZENG, 2005; HUANG; TZENG, 2006) foram usadas para encontrar padrões peculiares em situações onde as variáveis dependentes e independentes exibem relações complexas.

2.2.1 Aprendizado de Máquina em Risco de Crédito

O termo aprendizagem em máquina evoluiu a partir do estudo do reconhecimento de padrões e da teoria da aprendizagem computacional em inteligência artificial. Este campo explora o estudo e a construção de algoritmos que podem aprender e fazer previsões sobre dados previamente coletados (SAMUEL, 1959). Na gestão de risco de crédito, pode-se argumentar que cada conjunto de dados é peculiar e único em cada circunstância, portanto as relações dos dados podem ser bastante complexas, não-normais, não lineares e talvez não reflitam mudanças estruturais, como tendências demográficas ou de mercado. Portanto, a

construção e melhoria de tais modelos é dinâmica e um processo contínuo. (GALINDO; TAMAYO, 2000).

Quando um novo algoritmo é proposto para melhor e\ou prever situações de dificuldades financeiras de empresas, o primeiro passo é coletar um conjunto de dados, o qual é usado para criar um modelo que possa ser usado para generalizar novos dados recebidos. (ZHANG et al., 2003). Esse conjunto de dados é chamado de "conjunto de treinamento" e será usado para os algoritmos aprenderem com as características nele contidas.

O próximo passo envolve a preparação dos dados e, dependendo das características das informações coletadas, pesquisadores têm que enfrentar o desafio das “não-respostas” ou dados faltosos, escolhendo o melhor tratamento dentre muitas abordagens. (BATISTA; MONARD, 2003). Yu e Liu (2004) sugerem que a pesquisa não deve lidar com informações irrelevantes sobre o conjunto de dados. Portanto, a seleção de um subconjunto de variáveis torna-se pertinente, pois permite identificar e remover fatores redundantes ou desnecessários.

Markovitch e Rosenstein (2002) sugerem que a redução do número de dimensões dos dados aumenta a velocidade e a eficácia dos algoritmos na mineração de dados. O fato de muitas variáveis (dependerem uma das outras) serem interdependentes, geralmente influenciam a precisão dos modelos de classificação na aprendizagem da máquina. Os autores também recomendam que novas variáveis sejam formadas a partir do conjunto (de variáveis) das originais, contribuindo para a geração de novos classificadores que possam refletir melhor o conceito.

Mais especificamente, na modelagem de risco de crédito, um problema geral é encontrar variáveis, métodos e algoritmos adequados para permitir a instauração de modelos matemáticos ou estatísticos que possam ajudar a identificar padrões ou tendências de dificuldades financeiras. Recentemente, levando em consideração a complexidade dos dados produzidos pelo mercado, onde o ruído, a não linearidade e as idiossincrasias são a regra, a melhor estratégia lógica tem sido construir modelos que se beneficiam de uma abordagem interdisciplinar que combina estatísticas robustas e algoritmos de aprendizado de máquina. (FAYYAD et al., 1996; BIGUS, 1996; ADRIAANS; ZANTINGE, 1996).

Por essa razão, técnicas de inteligência artificial (IA), como Redes Neurais Artificiais (SMALZ; CONRAD, 1994; MALHOTRA; MALHOTRA, 2003; LAI et al., 2006); Algoritmo Genético (VARETTO, 1998; CHEN; HUANG, 2003) e *Support Vector Machine* (VAN GESTEL et al., 2003; HUANG et al., 2004), demonstraram resultados mais favoráveis - quando comparados a modelos estatísticos e técnicas de otimização para avaliação de risco de crédito - e vêm recebendo muita atenção no campo.

Embora quase todos os métodos possam ser usados para avaliar o risco de crédito, recentemente - devido à crescente complexidade e tamanho de base de dados - pesquisadores têm combinado diferentes classificadores e técnicas, que integram dois ou mais métodos de classificação e essas abordagens, têm mostrado maior precisão na previsibilidade do que qualquer método individual sozinho. A combinação de classificadores prosperou na avaliação de risco financeiro. Dentre alguns exemplos estão: Técnica Discriminante Neural (LEE et al., 2002), neuro-fuzzy (PIRAMUTHU, 1999; MALHOTRA; MALHOTRA, 2002) e fuzzy-SVM (WANG et al., 2005).

Os sistemas de combinação de técnicas são um exemplo de aprendizado de máquina onde vários classificadores são treinados para resolver o mesmo problema, e tentam construir um conjunto de hipóteses e as combinam. (LIU; ÖZSU, 2009.).

2.2 Metodologia

Designado a, primeiramente, vislumbrar o que pesquisa acadêmica alcançou na última acerca de ambos os tópicos, o Método ProKnow-C foi empregada. Uma ferramenta adotada por Ensslin et al. (2017) foca em uma análise de multicritérios construtivista na tomada de decisão (MCDA-C) quanto a escolha dos estudos. Uma metodologia que difere de outros métodos multicritério, principalmente reconhecendo os limites de conhecimento do pesquisador e organizando atividades em seu processo operacional que são consideradas decisivas e suficientes para avaliar o contexto do problema (DA ROSA et al., 2012; ENSSLIN et al., 2017).

Uma pesquisa científica começa com um problema que motiva o pesquisador a buscar informações sobre um determinado assunto em bases de dados bibliográficas (ENSSLIN et al., 2017). No entanto, as últimas décadas testemunharam um aumento espetacular de infraestruturas e recursos computacionais que são capazes de armazenar vasta quantidades de dados. Com esse desenvolvimento, o uso de bancos de dados como sistemas de indexação de periódicos, livros, teses, relatórios, anais de eventos, entre outros estudos, simplificou as buscas por referências bibliográficas e servem como plataforma teórica para futuras pesquisas.

Além de ser uma ferramenta para acelerar a busca e o uso do conhecimento científico em pesquisas, as bases de dados bibliográficas também contribuem com o estabelecimento de indicadores que são utilizados para avaliar o impacto e a relevância de um periódico em determinado ramo do conhecimento (GARFIELD, 2006). Uma vez determinado o campo do

conhecimento, o próximo passo é eleger as palavras-chave que guiarão a busca por referências (DA ROSA et al., 2012; ENSSLIN et al., 2017).

Para medir, interpretar e avaliar os resultados coletados, as pesquisas adotam técnicas bibliométricas que são análises quantitativas para avaliar publicações e disseminação científica (WANG; VEUGELERS; STEPHAN, 2017). A identificação do estágio atual do conhecimento em relação a esse nicho científico selecionado é considerada um aspecto crítico para que um acadêmico possa colocar seus objetivos de pesquisa dentro de um campo amplo e disperso de gerenciamento de risco de crédito. (TRANFIELD; DENYER; SMART, 2003).

2.3.1 A Base de Dados

A escolha das fontes de dados recebeu atenção especial devido às inúmeras opções de banco de dados online disponíveis em todo o mundo, que indexam publicações científicas e oferecem uma vasta gama de informações. No entanto, para o presente estudo, as seguintes bases de dados: *Web of Science, SCOPUS, Science Direct* e *Emerald* foram escolhidas para cobrir o maior número possível de publicações. Logo a escolha desses bancos de dados deveu-se à importância científica e abrangência de áreas que abrangem. Caso fossem encontrados artigos repetidos, eles seriam excluídos no decorrer do processo.

Uma vez determinado o campo do conhecimento, o próximo passo é eleger as palavras-chave que guiarão a busca por referências. Os autores submeteram a seguinte lógica booleana de palavras aos mecanismos de buscas dessas para delimitar o arcabouço conceitual deste estudo: (*OR Credit Risk OR Bankruptcy OR Credit Scoring OR Financial Distress OR Failure) AND (Machine Learning OR Data Mining OR Feature Selection OR Variable Selection OR Ensemble*).

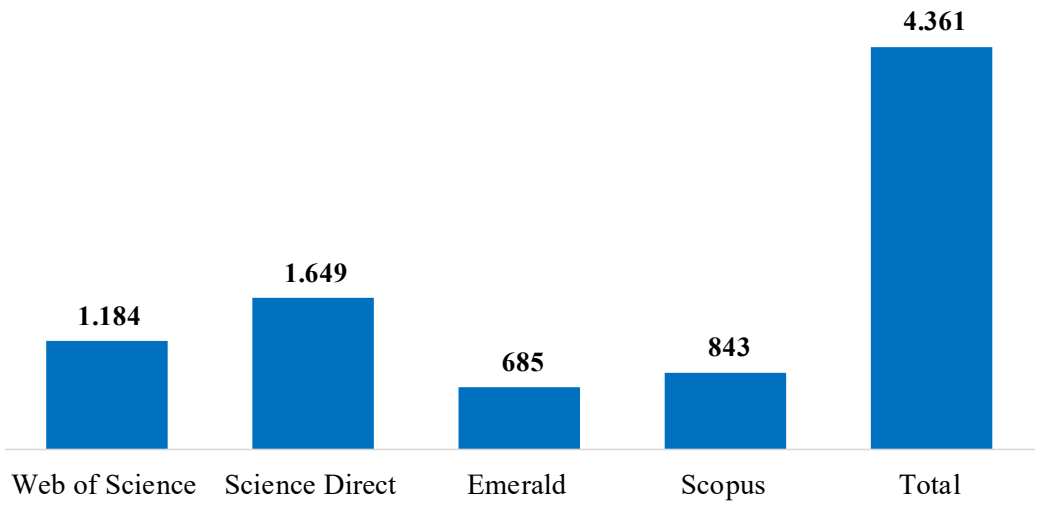
Após uma primeira investigação e leitura dos títulos dos primeiros artigos encontrados nas primeiras buscas, revelou-se eficaz o uso dessas palavras-chave para delimitar o objetivo deste estudo. Entretanto, com o intuito de explorar os artigos de um tamanho de (amostra) material representativo, a união desses dois blocos de palavras-chave foi usada para formar a amostra inicial de artigos que deu início ao processo de seleção para compor o referencial teórico.

2.3.2 Resultado das buscas

Usando as palavras-chave mencionadas anteriormente e delimitando a data de publicação entre janeiro de 2008 e setembro de 2017, a pesquisa gerou 4.361 resultados na

primeira amostra. Na figura 1, é possível ver como os resultados são distribuídos entre os diferentes bancos de dados.

Figura 1 – Número de registros encontrados correlacionados com as palavras-chave em cada base de dados.



Fonte: Resultados da pesquisa.

Desse primeiro total, foram excluídos 1.398 resultados, pois continham livros e capítulos de livros, permanecendo com uma primeira amostra de 2.963 referências. Então, desses 2.963 resultados foram excluídos 387 artigos duplicados, deixando o estudo com uma segunda amostra de 2.576 referências. Após a numeração e ordenação, os títulos de cada artigo foram lidos para observar o quão estavam alinhados com a pesquisa atual e, após essa investigação, 2.172 foram excluídos por não terem aderido ao tema principal da pesquisa, restando 404 artigos para posterior análise.

Essa amostra considerada alinhada com o objetivo da pesquisa foi então examinada com base em sua relevância científica e data de publicação. Foi então, dividida em dois grupos diferentes: o primeiro grupo compreendeu trabalhos publicados entre 01/01/2008 e 31/12/2015 e o segundo grupo entre 01/01/2016 e 30/09/2017. Essa divisão foi feita para evitar viés de seleção, quais sejam: (1) artigos recentes podem não ter tido tempo suficiente para serem reconhecidos cientificamente, mas poderiam ter potencial devido ao seu conteúdo recente; (2) estudos publicados recentemente tendem a ter menor número de citações e (3) não revisar apenas artigos cientificamente consolidados.

Com a intenção de padronizar o processo, buscou-se o número de citações de cada artigo no *Google* Acadêmico para verificar o número de citações. Com essa informação em (mãos) mão, estabeleceu-se um valor de corte de 90% para os artigos mais citados. Em outras palavras, 90% dos artigos mais citados e publicados antes de 2016, totalizavam 12.003 citações e os outros 10% correspondiam a 1.326 citações apenas. Ensslin *et al.* (2017) também executou esse processo. Em números absolutos, 90% das citações representava 165 artigos e 10%, 139 trabalhos. Assim, 90% dos artigos integram o repositório de estudos científicos com títulos e reconhecimento científico intitulado Repositório "k".

Em seguida, os 165 artigos que representaram mais de 90 % da relevância científica foram analisados quanto ao alinhamento de seus resumos. Nestes, seus 158 resumos foram julgados consistentes com o tópico de pesquisa e 7 foram excluídos. Os autores desses 158 artigos compuseram o "banco de dados de autores", que consistiam de 300 pesquisadores.

Os 10% restantes dos artigos foram colocados em um repositório diferente, denominado Repositório "n", e classificados de acordo com o alinhamento do título, mas com menor relevância científica. Na tentativa de evitar vieses de seleção, 100 dos resultados mais recentes - menos de 2 anos desde a data de publicação, entre 01/01/2016 e 30/09/2017 - foram agrupados com os 10% dos artigos menos relevantes para uma inspeção mais detalhada, totalizando 239 referências no Repositório "n".

Deste Repositório "n" leu-se os resumos dos estudos mais recentes, cuja data de publicação foi após 31/12/2015. Essa análise evitou a eliminação de artigos relevantes para esta pesquisa. Para os 10% dos artigos do repositório "n", somado aos estudos com menos de 2 anos desde sua publicação, verificou se algum dos autores fazia parte do banco de dados de autores construído a partir dos artigos com relevância científica já confirmados. Dessa maneira, foram mantidos 10 artigos e 229 foram excluídos - devido à sua falta de citação, relevância dos autores e alinhamento com esta pesquisa.

Nessa primeira análise, pode-se notar que autores responsáveis por estudos mais relevantes, publicaram menos nos últimos anos e a maioria dos pesquisadores ainda não conseguiram estabelecer uma relevância cientifica de suas pesquisas.

Neste momento, uma prévia com 168 artigos diferentes de ambos repositórios - cujos resumos foram analisados quanto ao alinhamento com o tema de pesquisa de ambos os repositórios - compreendem o referencial teórico final para este estudo. O último passo na construção do corpus desta pesquisa consiste em unir os artigos cuja relevância científica tenha sido verificada (Repositório "a") com os artigos menos relevantes e estudos recentes de (Repositório "n").

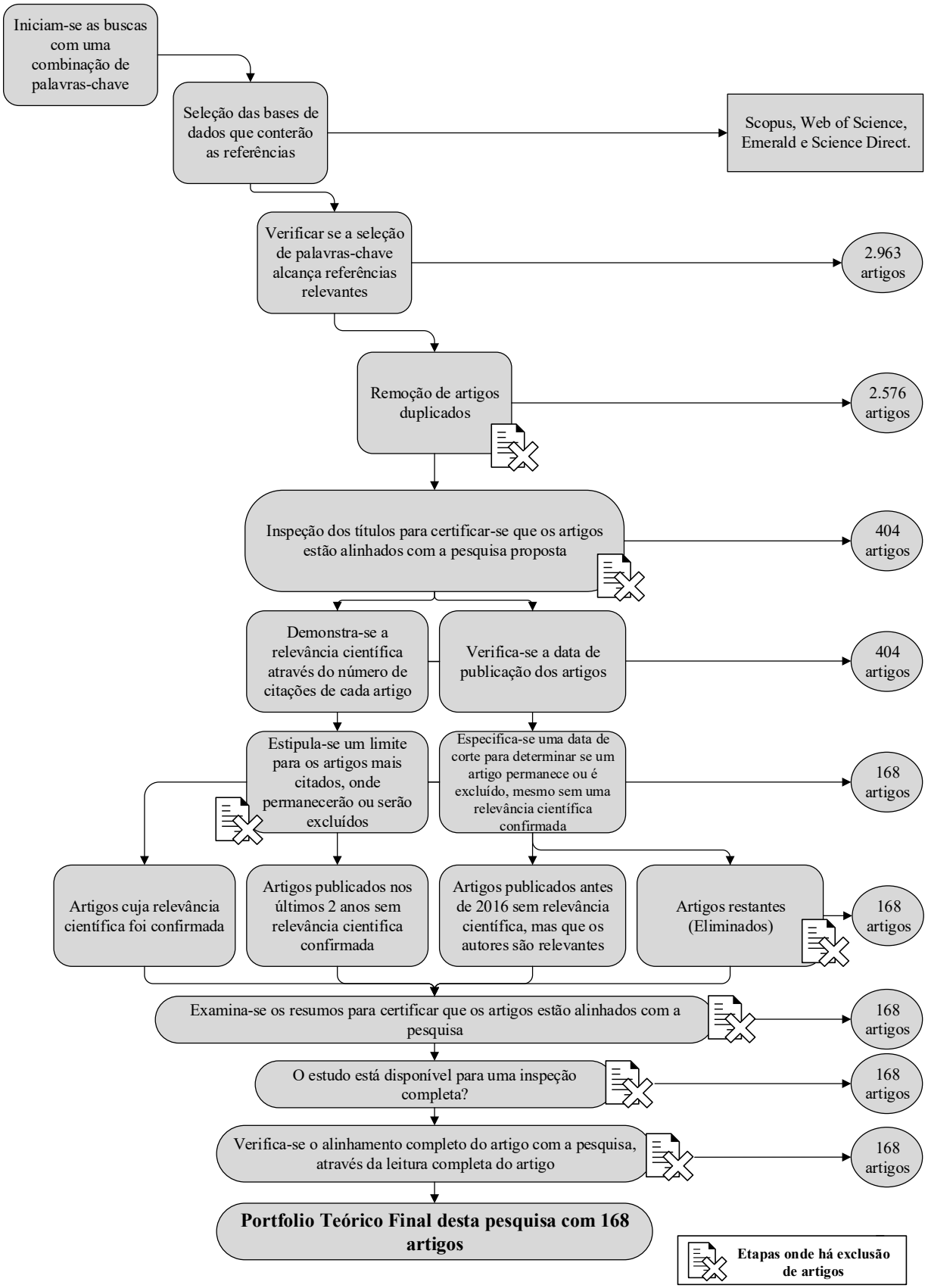
Por fim, buscou-se a disponibilidade de artigos para consulta (até o momento foram trabalhados apenas os títulos, citações e resumos). Caso o artigo não estivesse disponível na íntegra, ele seria descartado do portfólio final. Para aqueles artigos que (estão) estavam disponíveis na íntegra, procedeu-se a leitura completa e, finalmente, o alinhamento com o tema da pesquisa (é) foi revisado. Reiterando, aqueles artigos considerados alinhados, permanecem no banco de dados da pesquisa e acaba-se, então, a composição do referencial teórico da atual pesquisa. Foi possível encontrar uma cópia de cada artigo na instituição que apoia esse estudo, portanto, não foi necessário excluir nenhum outro artigo. Após a conclusão de todos os passos necessários, este estudo fundamenta-se em um portfólio bibliográfico final de 168 artigos.

Vale ressaltar que essa amostra bibliográfica não representa o estado da arte primorosa quando se trata de pesquisa em cada campo, mas sim a intersecção entre eles, consideradas as delimitações da pesquisa. Ou seja, caso outros pesquisadores repliquem o esse método, o portfólio bibliográfico manterá os artigos principais; alguns podem não ser selecionados, mas outros, não incluídos aqui, poderem ser contemplados.

O emprego do processo no contexto de pesquisa segue todos os passos descritos na seção anterior e representados na Figura 2 de forma ilustrativa. Como parte subsequente da construção do portfólio bibliográfico para o tema de pesquisa, está a análise quantitativa dos artigos selecionados em cada etapa para compor o portfólio final. Tal análise, tem o propósito evidenciar informações pertinentes ao tema de pesquisa por meio da análise e quantificação de suas características (DA ROSA et al., 2012; ENSSLIN et al., 2017). Ainda na figura 2, é possível verificar o resultado da seleção de artigos em cada etapa do processo. Nos próximos parágrafos, serão descritas as análises quantitativas do portfólio final e as decisões tomadas a partir dos resultados encontrados na leitura do portfólio final na aplicação da metodologia para atingir os objetivos do trabalho.

Os 168 artigos selecionados pelo processo podem ser encontrados no Apêndice I, (no) ao final da pesquisa.

Figura 2 - Processo de seleção de artigos por meio do método Proknow-C, similar a Da Rosa, 2012; Ensslin et al., 2017

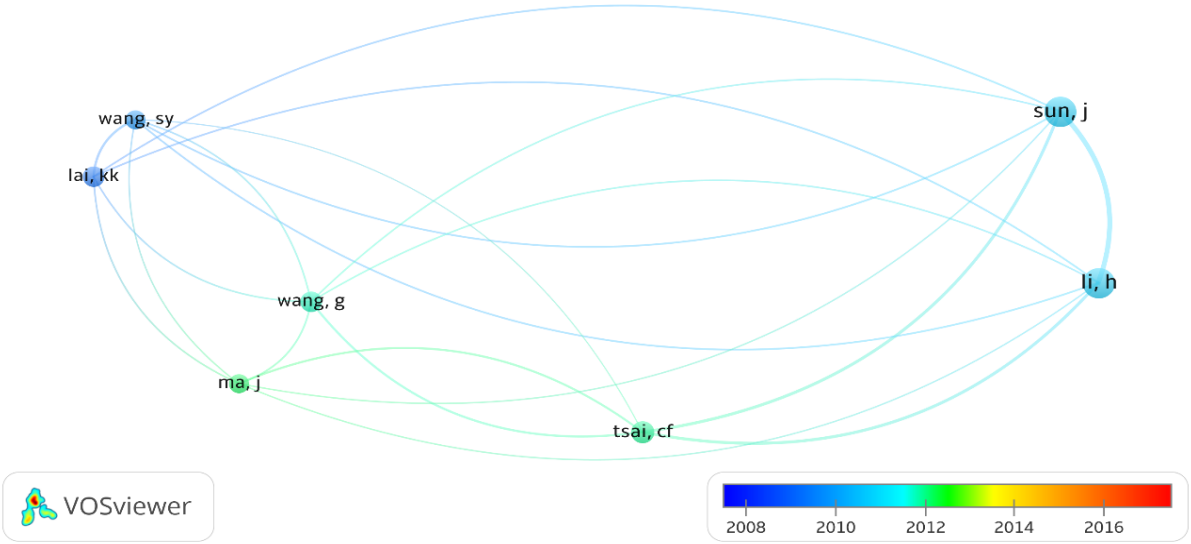


Fonte: Elaborada pelos autores.

2.3 Análise Quantitativa

O portfólio final englobou um total de 431 pesquisadores diferentes; mas, entre eles, apenas 60 autores publicaram mais de um artigo no intervalo de tempo selecionado sobre o assunto em questão. Hui Li e Jie Sun são os autores mais populares e cada um deles teve exatamente 20 trabalhos (cada) no portfólio final. O terceiro autor mais popular é Chih-Fong Tsai, com 8 publicações. Quando se eleva o número de publicações para 5, o total de pesquisas cai significativamente e pode-se ver que apenas 7 autores atendem a esse quesito.

Figura 3 - Autores mais ativos



Fonte: VOSviewer – Resultados da pesquisa.

Na figura 3 Li, H.; Sun, J; Tsai, C.F.; Wang, G; Lai, K.K.; Wang, S.Y. e Ma, J. como pesquisadores mais ativos nessa área de pesquisa. Para medir a relevância destes, o número de publicações, o número de citações e o índice-h foram envolvidos aqui para avaliar a qualidade profissional desses autores.

Tabela 1 - Relevância dos autores nesta pesquisa.

Autores	N° de publicações	N° de citações	Índice h
Li, H.	20	516	62
Sun, J.	20	516	59
Tsai, C.F.	8	419	21
Wang, G.	6	283	56
Lai, K.K.	5	362	41
Wang, S.Y.	5	301	43
Ma, J.	5	148	25

Fonte: Resultados da pesquisa.

Um dos autores de destaque, Hui Li, é professor da Universidade Nankai na China e, entre suas especialidades, ele se concentra em Raciocínio Baseados em Casos e Previsão, mineração de dados, previsão de falências, auxílio à tomada de decisões financeiras e sistemas de informações contábeis baseado em Aprendizagem de Máquinas. Em um dos artigos selecionados - dentro deste portfólio - onde ele é coautor, propõem duas novas abordagens dinâmicas de previsão de dificuldade financeira baseadas em "*Time Weighting"* e conjuntos estatísticos de *Adaboost-Support Vector Machine*. (SUN et al., 2017). A abordagem foi a de utilizar vários conjuntos estatísticos, unidos na forma de algoritmos de aprendizagem para, juntos, obterem melhores desempenhos preditivos.

Outro pesquisador que se destaca é Jie Sun. Hui Li e Jie Sun, como observado em toda a pesquisa, estão frequentemente trabalhando juntos. Sua linha de pesquisa seguem as mesmas de Hui Li. Contudo, nos artigos selecionados nesta pesquisa - aqueles em que ela é a principal autora – observou-se que o Raciocínio Baseado em Casos (CBR) e seleção de variáveis é sua área de maior foco.

Acrescentando à lista de pesquisadores, Chih-Fong Tsai fica atrás dos dois autores anteriores. Apesar de Chih-Fong Tsai não se concentrar particularmente a estudos de finanças, seus artigos selecionados neste estudo destacam o desenvolvimento de modelos usando *Soft Classification Techniques*, os quais incluem técnicas que envolvem conjuntos de classificadores e classificadores híbridos. O pré-processamento de dados para a seleção de variáveis é outro tópico muito utilizado pelo autor.

No portfólio final desta pesquisa, Shouyang Wang e Kin Keung Lai trabalharam juntos em 5 dos artigos selecionados. Dentro da área de gerenciamento de risco de crédito, ambos os autores optaram por usar os modelos de *Support Vector Machine* nos estudos mais

recentes. Shouyang Wang tem interesses em pesquisa com métodos de previsão e análise econômica por meio de programação matemática.

Gang Wang é outro pesquisador relevante que avoluma a relação. Ele incorpora aprendizado de máquina (algoritmos de *boosting* e *clustering*) para maximizar a precisão; seu outro foco é com a seleção de variáveis. Em Wang, Ma e Yang (2014), ele optou pela estratégia de seleção de variáveis juntamente com *Boosting* e sugeriu que o modelo obteve um melhor desempenho como base de aprendizagem.

Por último, há Jian Mac, co-autor em alguns estudos com Gang Wang. Em seu artigo com maior número de citaçoes nesta pesquisa, ele propõe um modelo construído baseado em Árvores de Decisão com duas estratégias de agrupamento: *bagging* e *random space*, cujos resultados sugerem que esses modelos poderiam ser utilizados como técnicas alternativas de análise de crédito. Em outro estudo, ele e outros pesquisadores sugerem que uma limitação dos métodos de aprendizagem estatísticas é a falta de interpretabilidade dos resultados. Eles admitem que as respostas dadas pelos modelos são de difícil compreensão por seres humanos e sugerem que melhorar a compreensão dos resultados dos modelos de aprendizagem é uma importante lacuna de pesquisa ainda pouco estudada.

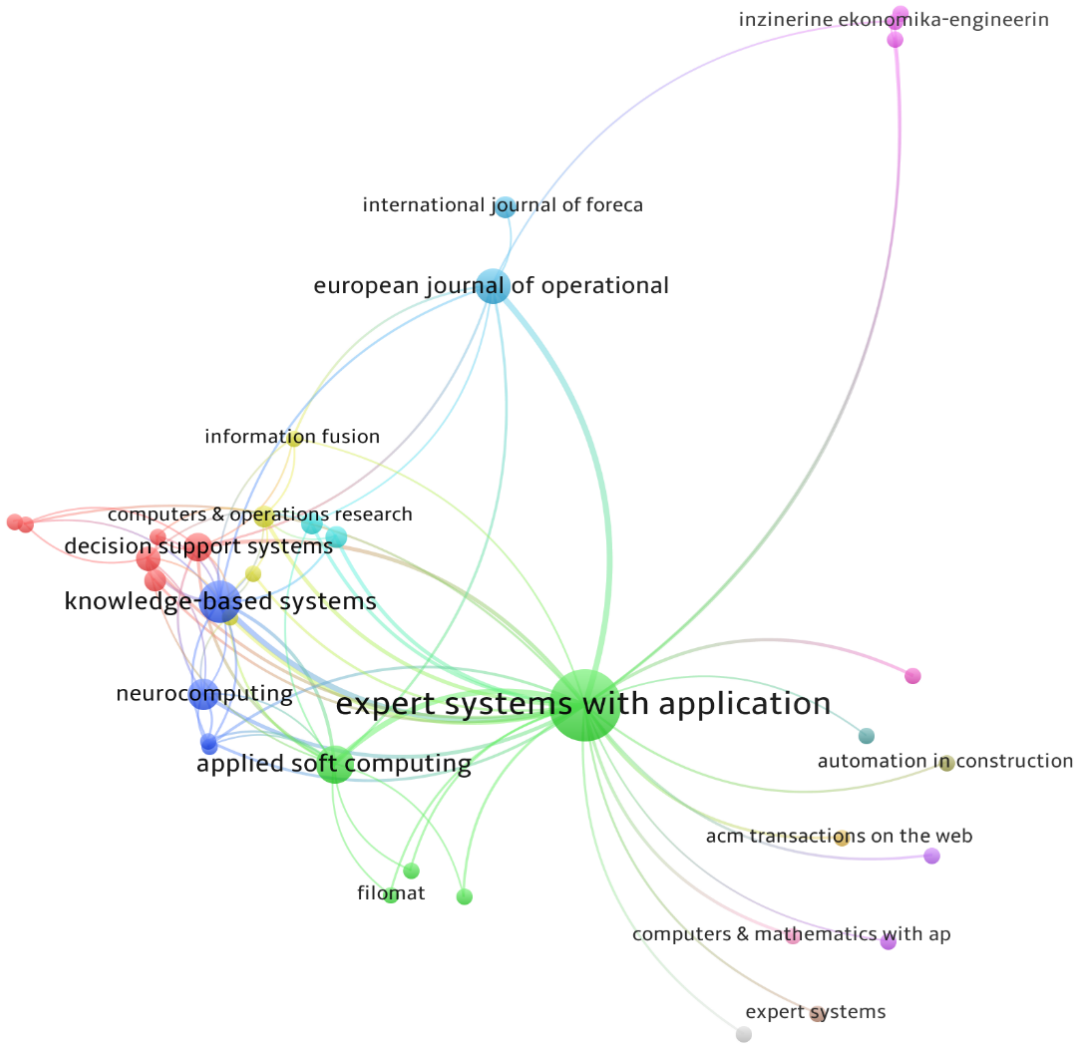
Esta breve revisão dos principais estudos dentro desta pesquisa demonstra que a Escola Chinesa lidera a corrida no campo da gestão de risco de crédito, contudo esses pesquisadores estão atuando não só em áreas de gestão de negócios, mas também nas áreas da ciência da computação. Pode-se concluir, com base na análise até o momento, que os estudos mais citados não são necessariamente focados em um tipo de técnica. Eles se aventuram por todo o campo experimentando e combinando métodos para cada etapa do processo e a China tem uma vantagem sobre o avanço do campo usando o aprendizado de máquina.

2.4 Artigos de acordo com sua relevância científica

Esta análise mostrou que os artigos do portfólio foram publicados em 43 periódicos diferentes. Entre eles, o *Expert Systems with Applications*, com 79 publicações, é a principal fonte. O próximo de maior destaque, com 15 publicações, foi *Knowledge Based Systems*. A partir dessa análise, pode-se observar que as publicações nos periódicos mais bem avaliados, dentro desse nicho de pesquisa, estão voltadas para a concepção e teste de modelos práticos de risco de crédito como ferramenta na tomada de decisões. Mesmo com avanços identificados nessas literaturas, não foi mencionado a substituição do papel de pessoas nos processos decisórios, por máquinas.

A figura 4 foi elaborada a partir dos resultados encontrados nas análises dos artigos e, a partir destes, pode-se mapear os periódicos de maior peso e como se relacionam entre si.

Figura 4 - *Journals* de maior destaque na pesquisa



Fonte: VOSviewer – Resultados da pesquisa.

Para medir a relevância dos periódicos no portfólio bibliográfico final, foram considerados os indicadores de classificação JCR (THOMSON REUTERS, 2017) e SJR (SCIMAGO, 2017). A Tabela 2 traz a posição dos 10 principais periódicos de acordo com seu fator de impacto dentro desta análise.

Tabela 2 - Medida de relevância científica dos Journals.

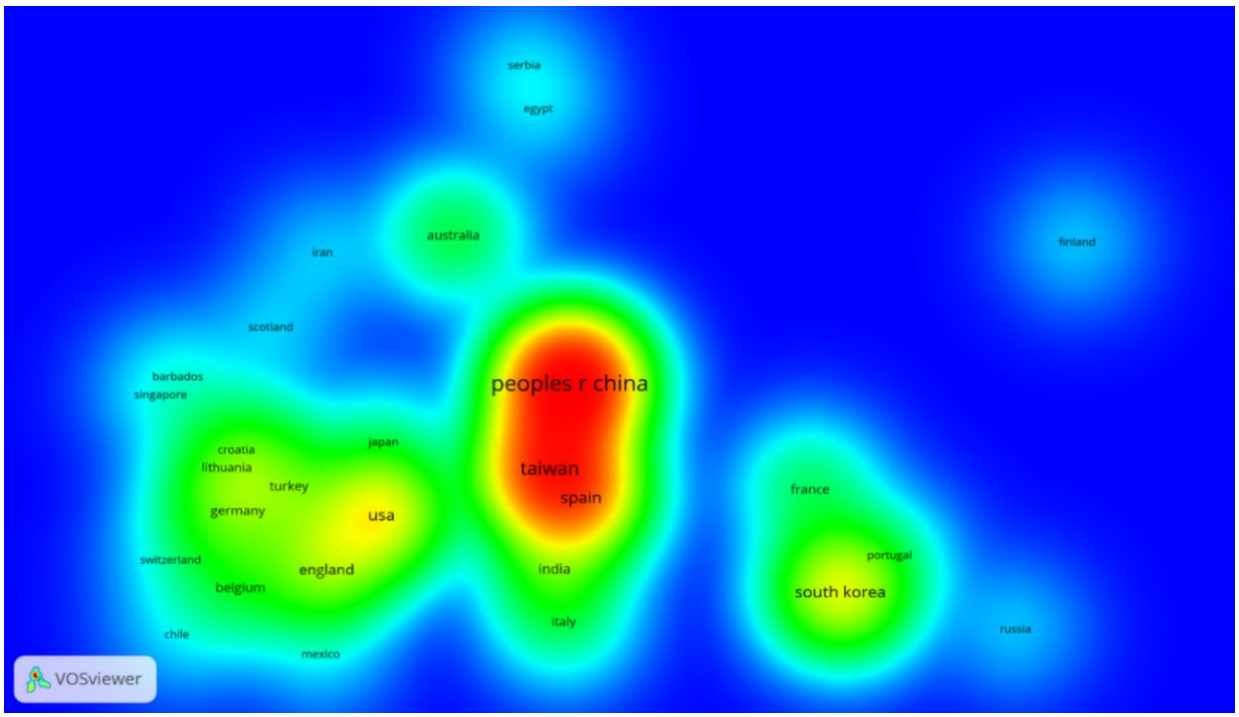
Journal	JCR	SJR
Expert Systems with Applications	2981	1,433
Knowledge-Based Systems	4627	1,877
Applied Soft Computing	3811	1,308
European Journal of Operational Research	3526	2,505
Neurocomputing	3317	0,968
Decision Support Systems	4290	1,806
Journal of Forecasting	2837	1,682
International Journal of Neural Systems	6333	1,121
Information Sciences	4732	1,91
Journal of Banking & Finance	2570	1,767

Fonte: Resultados da pesquisa.

Agora, por meio de uma análise dos países de origem dos autores de maior destaque nessa pesquisa, pode-se elaborar um mapa de procedência desses estudos. Na figura 5, é possível ver que os autores asiáticos publicaram a maioria dos artigos de 2008 a 2017 e o maior número de citações são encontradas na China e em Taiwan, com 3.034 citações. Ao examinar o mapa de densidade, pode-se perceber que as regiões externas do mapa mostram países onde as pesquisas de risco de crédito são escassas.

É importante notar, como mostra a figura 5, que estudos realizados na China e Taiwan servem como referência para vários outros realizados na Europa, América do Norte e Oceania. Ademais, é curioso ver que regiões como a América Latina e África ainda não foram capazes de produzir trabalhos com relevância científica. Estudos contemplando informações dessas regiões devem produzir resultados interessantes em futuras pesquisas.

Figura 5 - Países por número de publicações



Fonte: VOSviewer – Resultados da pesquisa.

2.5 Análise quantitativa de artigos selecionados

O primeiro passo para analisar as pesquisas abordadas nos artigos do portfólio final foi codificá-las de acordo com as categorias descritas na tabela 3 para análise quantitativa dos dados encontrados. São elas: (1) Assunto Principal, (2) Método, (3) Tipo de Fonte de Dados, (4) Região Geográfica do Estudo, (5) Variável de Interesse e (6) Variáveis Independentes. Após a classificação, essas categorias serão analisadas separadamente.

Tabela 3 - Classificação utilizada para analisar os artigos.

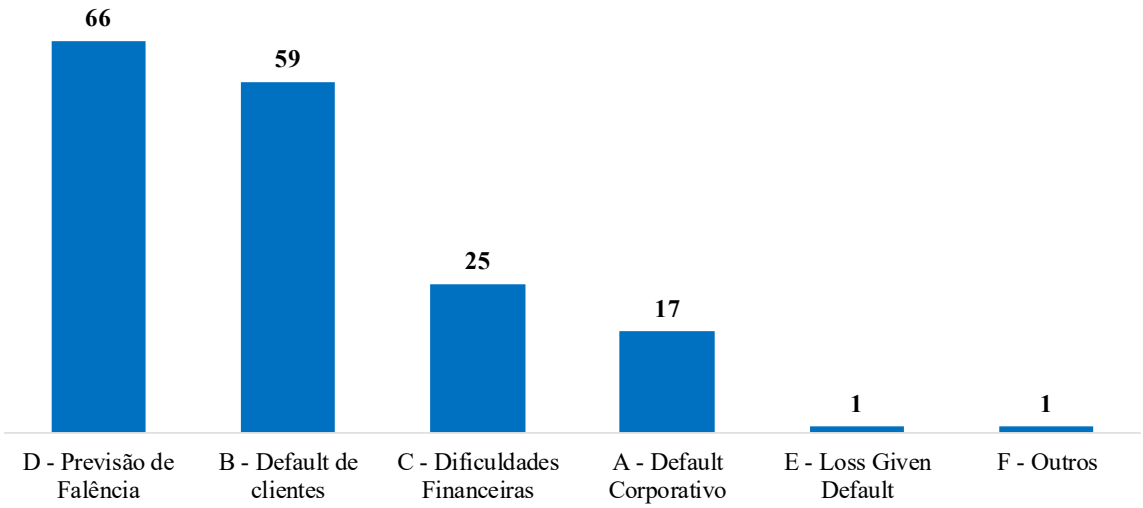
Categorias	Classificação
Assunto Principal	A - *Default* Corporativo B - *Default* de clientes C - Dificuldades Financeiras D - Previsão de Falência E - *Loss Given Default* F - Outros
Método	A - Métodos Estatísticos B - Árvore de Decisão C - *Support Vector Machine* D - Rede Neural Artificial E - *K-Means* F - Seleção de Variáveis G - Algorítmos de *Gradient Boostin* H - Combinação de Técnicas
Tipos de Fontes de Dados	A - Instituições Financeiras B - Intituições Não-Financeiras C - Outros
Fonte Geográfica de Dados	A - América do Norte B - Europa C - Asia D - Mercados Emergentes E - Outros / Não Mencionados (Australia)
Variável de Interesse	A - Falência B - Dificuldade (Financeira) C - *Default* D - *LGD*
Variáveis Independentes	A - Índices Financeiros B - Índices Não-Financeiro C - Informações Pessoais D - *LGD (Loss Given default)*

Fonte: Elaborado pelo autor.

Ao analisar o tema risco de crédito, foi possível concluir que as pesquisas estão focadas na previsão de falência e dificuldades financeiras, seguida pela estimação de escore de crédito do consumidor. Na figura 6, pode-se observar que diferentes aplicações de gestão de risco de crédito devem ser examinadas mais detalhadamente. Coudert e Gex (2010) buscaram entender se um efeito de contágio teria sido causado pela General Motors e Ford em maio de 2005 em todo o mercado americano de *Credit Default Swap* (CDS). Em suas

análises, eles sugeriram que as correlações entre o CDS e CDS emitidos pela GM e Ford aumentaram significativamente (cerca de 17%) durante a crise de 2005. A partir desta análise levanta-se a seguinte questão: em uma economia onde há forte integração dos mercados, indústrias específicas poderiam desencadear uma crise de crédito?

Figura 6 - Classificação dos Assuntos Principais.

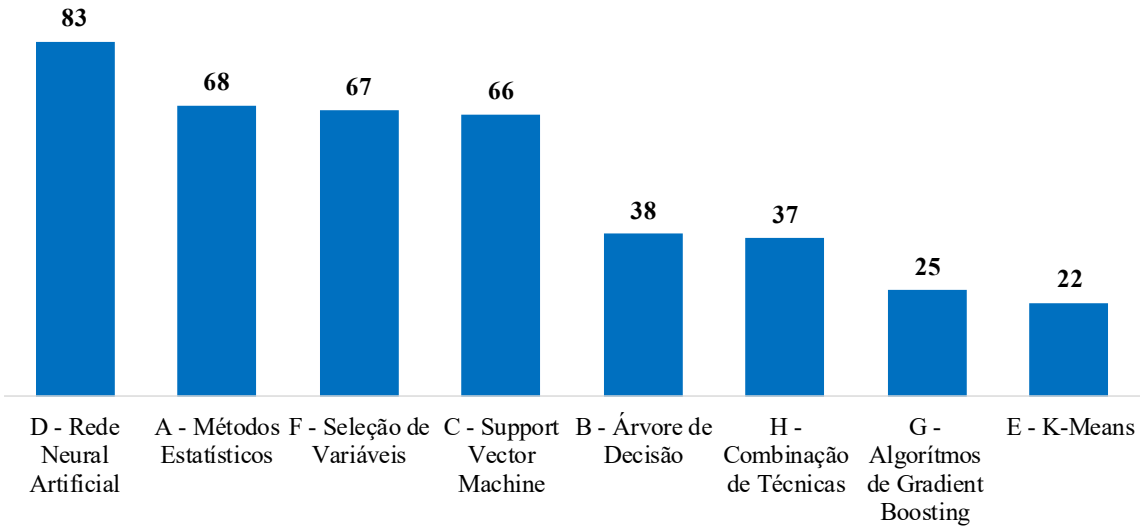


Fonte: Elaboração própria

Verificou-se que apenas Loterman *et al.* (2012) analisou o risco de *Loss Given Default (LGD)*. Em sua descoberta, eles não tiveram um desempenho significativo dos modelos, sugerindo que as aplicações de técnicas não-lineares à modelagem de LGD poderiam melhorar o desempenho e a compreensibilidade dos modelos. Por via dessa pesquisa, constatou-se que pesquisadores de risco de crédito geralmente focam na probabilidade de inadimplência, contudo, o parâmetro LGD (que mede a perda financeira, expressa como percentual da exposição total, em caso de inadimplência) não foi analisado com mais rigor ou de forma mais ampla.

A partir desses resultados, presume-se que há necessidade de se concentrar na recuperação financeira. Ou seja, uma vez que o cliente tenha deixado de amortizar as parcelas devidas, qual é o esforço necessário para recuperar total ou parcialmente o crédito cedido e como medi-lo. Nesta fase da análise quantitativa, não foi possível identificar por que os estudos sobre o efeito de contágio, LGD, ou até mesmo *Exposure at Default* (EAD), não foram receberam atenção ou não foram considerados, evidenciando assim, mais uma lacuna para futuras pesquisas.

Figura 7 - Classificação dos principais métodos utilizados.



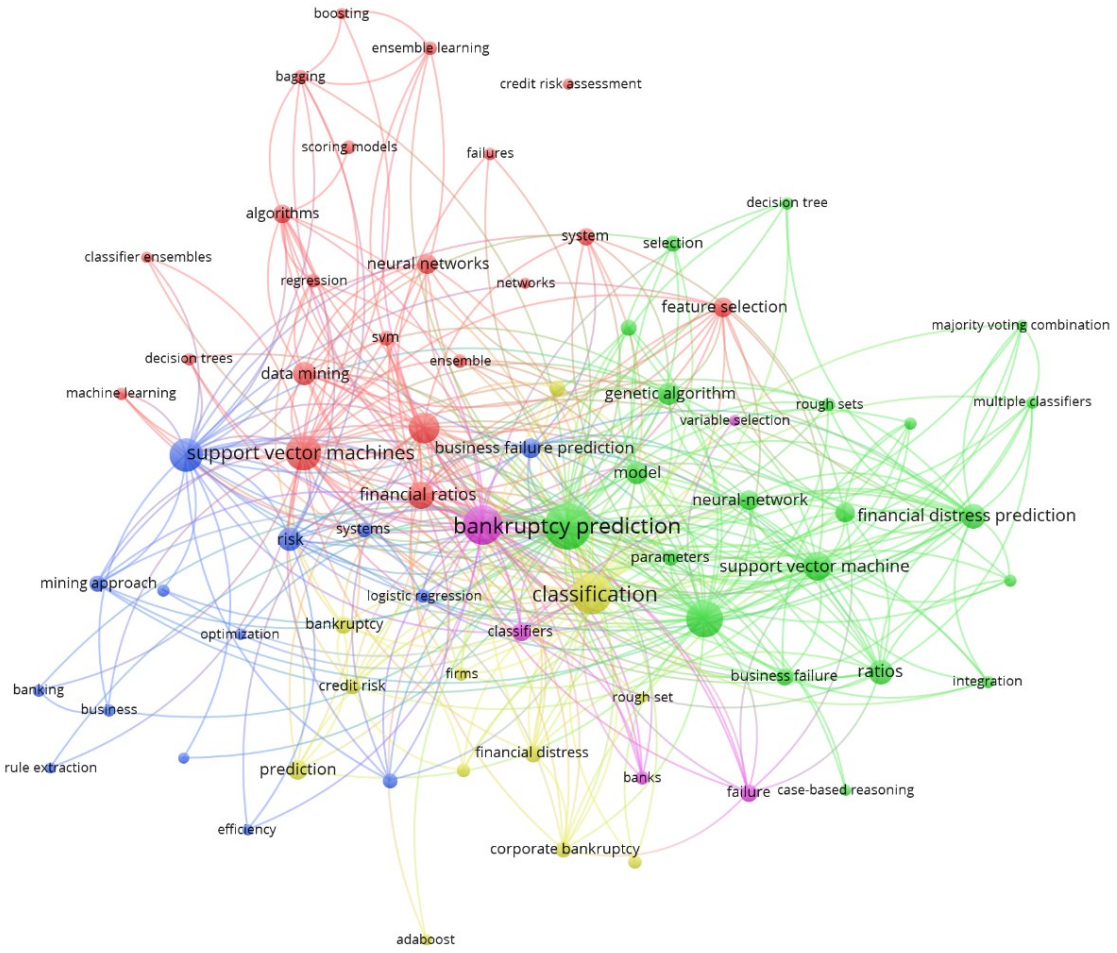
Fonte: Resultados da pesquisa.

Avançando na análise, a figura 7 mostra os modelos utilizados e pode-se observar que *K-means* foi a última escolha como método, a partir desta constatação, pode-se inferir que algoritmos que não são capazes de lidar com dados que contém muitos ruídos e *outliers,* são considerados ineficientes para dados não-lineares. Devido a esta condição de não-linearidade das informações, algoritmos que convertem a capacidade de classificadores-bases de aprendizagem em classificadores mais fortes, mostram ser capazes de produzir resultados significativos e expõem mais lacunas a serem exploradas. O uso de Redes Neurais aparece como a opção de preferência pelos pesquisados, seguidas por análises estatísticas tradicionais.

Outro método bastante utilizado foi *Random Forest* e os resultados encontrados pelas pesquisas apontam para sua eficiência e precisão, dado que o algoritmo cria várias árvores de decisão e as combina para obter um prognóstico com maior acurácia e mais estável.

Um caso relevante de acordo com Wang e Ma (2011) e seu método é que a melhoria na interpretabilidade dos resultados é outra importante motivação de pesquisa ainda pouco estudada. Além disso, Zhang et al. (2010) sugerem que uma melhor maneira de aprimorar a precisão da classificação seria utilizar estratégias que combinem resultados de vários modelos individuais em vez de construir um único modelo de alta capacidade preditiva.

Figura 8 - Mapa de Palavras-chave.



Fonte: VOSviewer – Resultados da pesquisa.

Também foi possível identificar lacunas na literatura quando se considera os níveis mais externos da figura 8. Ao tentar prever a variável de dependente usando qualquer técnica de aprendizado de máquina, a escolha por um conjunto de técnicas que se complementam, têm sido uma escolha gradual por sua natureza de classificação ao fornecer uma previsão final. Um dos motivos pelos quais pode-se notar que a combinação de técnicas vem ganhando espaço, como alternativa, é a simples ideia de que diferentes modelos preditivos que tentam prever a mesma variável dependente podem desempenhar uma tarefa melhor do que qualquer método isolado.

2.6 Métricas de Performance

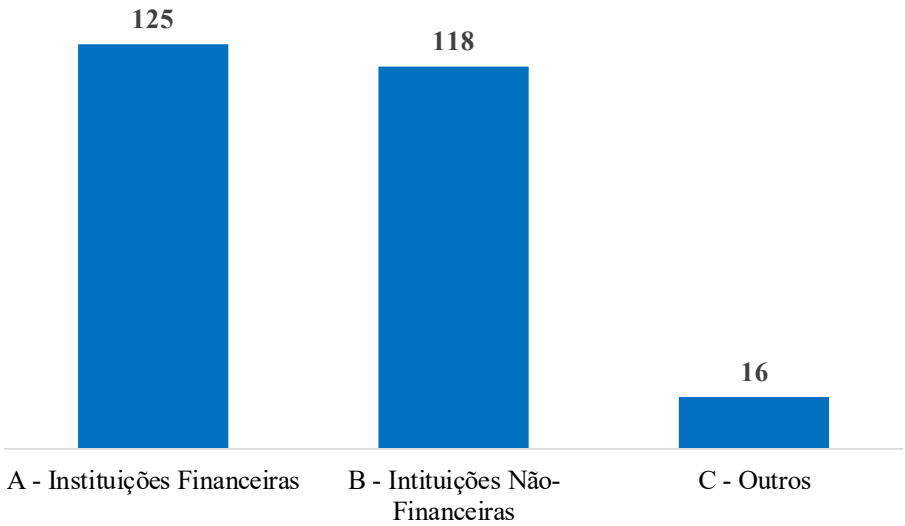
O desempenho dos modelos é geralmente avaliado em termos de algumas métricas, dentre as quais, têm-se: precisão (ACC), erro quadrático médio (MSE), taxa verdadeiros positivos (TPR), taxa verdadeiros negativos (TNR), Medida-F, área sob curva ROC (AUC).

Li e Sun (2009) sugerem em seu estudo que o modelo proposto alcançou um desempenho preditivo aceitável mesmo em uma condição onde não tinham nenhuma evidência sobre quais características são relevantes para o objetivo pesquisado e quais características são irrelevantes. Portanto, essas métricas podem avaliar o desempenho geral do modelo, entretanto, deveriam ser mais descritivas ao informar quais variáveis estão influenciando o desempenho dessas métricas.

2.7 Variáveis dependentes

Os modelos desenvolvidos e abordados nos artigos desta análise dependiam principalmente de índices financeiros e variáveis contendo informações pessoais. Algumas fontes de dados não foram informadas, portanto foram classificadas como "Outros". A partir de uma análise dos conjuntos de dados, pode-se perceber que as bases de dados disponíveis publicamente têm sido usadas sistematicamente, o que pode levantar questões sobre a replicação de tais modelos em dados de empresas ou bancos, embora essas bases também contenham informações reais.

Figura 9 - Classificação das principais fontes de dados estudadas.

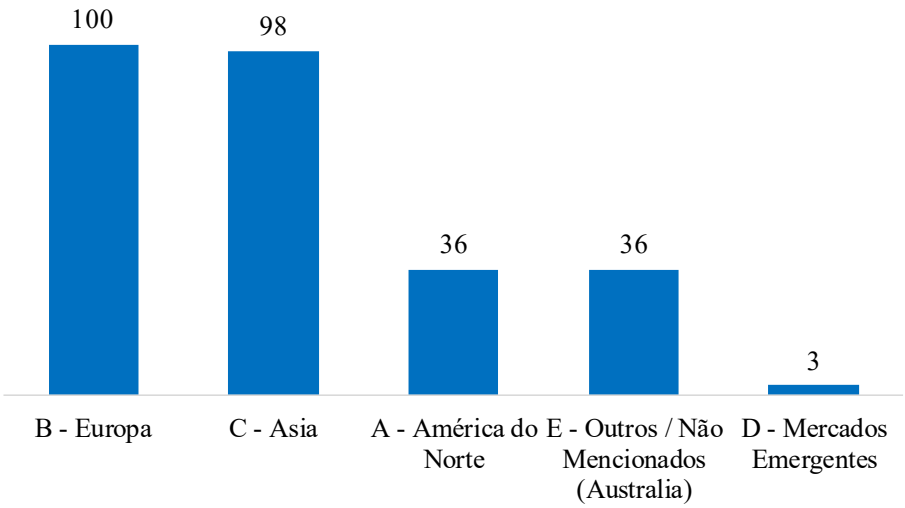


Fonte: Resultados da pesquisa.

Quanto a origem desses dados, na figura 10 pode-se notar que se tem utilizado dados, principalmente europeus, para construir modelos. Esses dados pertinentes a clientes foram disponibilizados por informações de instituições financeiras. Logo um maior número de estudos está concentrado em análise de escores de credito de pessoas.

Uma observação interessante sobre a categoria de dados utilizados foi feita por Guo *et al* (2016b), no qual os autores utilizaram dados sociais para conduzir o estudo. Nesta pesquisa, ambos destacam a crescente popularidade das redes sociais, como o Twitter e o Weibo, e que indícios comportamentais de usuários estão se acumulando rapidamente; os quais poderiam conter sinais correlacionados ao uso de crédito embutido nestas informações. Diante disso, sugerem que futuras pesquisas concentrem-se em desenvolver métodos mais eficazes de mineração desses dados e a incorporação das variáveis para melhor entender as possibilidades e limites desse tipo de informação na construção de *escores* de crédito.

Figura 10 - Classificação geográfica das fontes de dados.

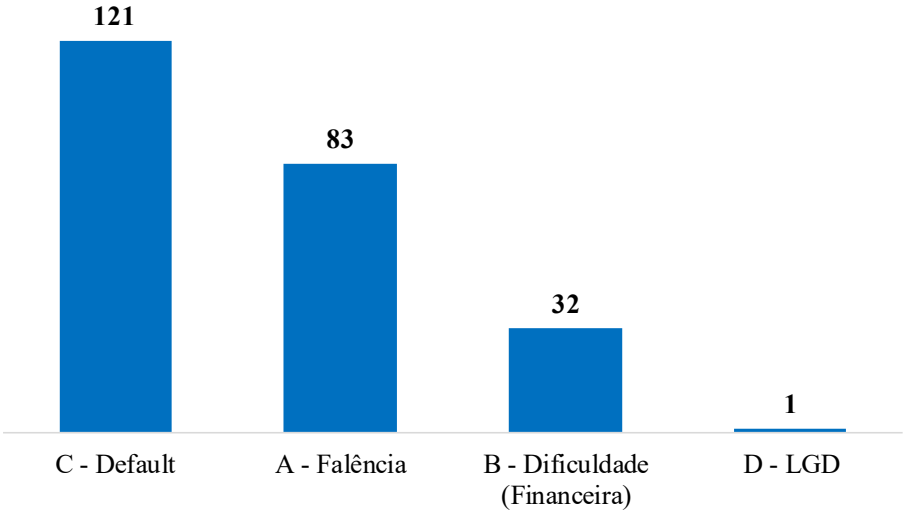


Fonte: Resultados da pesquisa.

Todavia pouca atenção continua sendo dada aos mercados emergentes. Nenhum dos acadêmicos mais influentes utilizou informações financeiras dessas regiões, tampouco pesquisadores dessas regiões emergentes conseguiram produzir artigos com relevância científica. Em relação à disponibilidade de informações corporativas ou de consumidores dessas regiões, não se pode verificar se os pesquisadores tinham acesso a esses dados.

A partir das bases de dados utilizadas, as variáveis dependentes foram agrupadas conforme a figura 11, e nota-se que estudos, são primeiramente preocupados com a probabilidade de inadimplência, um parâmetro crucial nas solicitações de empréstimos, estimativa de *rating*, precificação de derivativos de crédito e muitos outros aspectos em finanças.

Figura 11 - Classificação das variáveis de interesse.

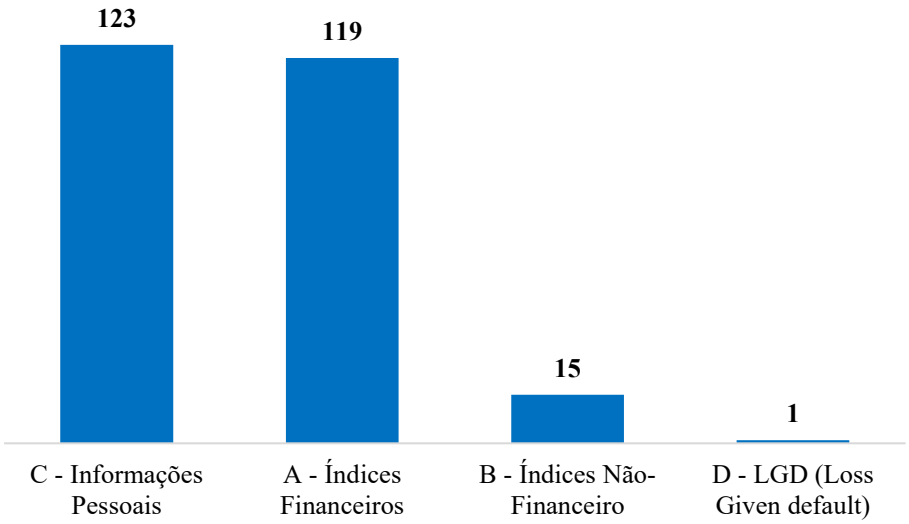


Fonte: Resultados da pesquisa.

Pesquisadores e empresas de crédito tentam minimizar qualquer eventual perda financeira, de modo que a eficiência da administração de uma corporação é reconhecida como um fator que colabora para falência de empresas, no entanto é geralmente excluída de modelos buscam antecipar dificuldades financeiras (YEH; CHI; HSU, 2010). Isso indica outro aspecto importante da pesquisa que deve ser aprofundado: a adoção de diferentes variáveis qualitativas. A partir da figura 12, nota-se que a maioria das variáveis independentes usadas nos modelos representam aspectos financeiros.

Liang *et al.* (2016) sugerem em seu artigo que modelos combinando índices financeiros e indicadores de governança corporativa podem melhorar o desempenho do modelo quando comparado com o modelo baseado apenas em índices financeiros. Posto a complexidade de capturar eventos macroeconômicos ou vicissitudes pessoais por meio de dados numéricos, sugere-se então a formulação de bases de dados que contenham variáveis qualitativas que possam indicar ou minimizar o impacto de momentos econômicos na análise de crédito.

Figura 12 - Classificação de variáveis independentes.



Fonte: Resultados da pesquisa.

2.8 Considerações finais

A previsão e gestão de risco de crédito é uma questão exploratória que está em alta nas áreas de contabilidade, finanças, e agora, computação. Ela foi gradualmente construindo seu próprio corpo teórico para ser um assunto independente, pois tem implicações práticas importante na compreensão e melhoria do risco financeiro, evitando dificuldades financeiras, falências e mensurando a probabilidade de inadimplência de um determinado grupo ou perfil no qual um consumidor se insere. Essa área de estudo ganhou nova direção quando a aprendizagem de máquina foi introduzida. Seu intuito é o uso da inteligência artificial (IA) em sistemas cuja capacidade de aprender e melhorar automaticamente a partir da experiência, não tenha que ser explicitamente programado ou reajustado. As descobertas deste estudo sugerem que as previsões feitas por máquinas inteligentes, são consideravelmente mais adaptáveis e capazes de captar a dinâmica dos ciclos de crédito.

Esta pesquisa buscou coletar e selecionar as publicações científicas mais relevantes, desde a última crise em 2008, sobre o risco de crédito sob a ótica da aprendizagem de máquina. A intenção foi a de analisar qual direção esta área de pesquisa está seguindo e identificar lacunas na literatura que devem ser examinadas mais detalhadamente. Por esse motivo, foi proposto o uso da metodologia de pesquisa ProKnow-C para atingir esse objetivo.

A análise quantitativa desses artigos promoveu um processo de aprendizagem - por meio de uma abordagem construtivista - que auxiliou os pesquisadores deste estudo a

compreender e definir os periódicos, artigos e autores relacionados ao tema de pesquisa. A análise bibliométrica dos 168 artigos possibilitou a obtenção de achados interessantes e, embora numerosos artigos tenham sido publicados neste campo, ainda existem alguns tópicos valiosos que precisam ser melhor explorados e estudados.

Ao rever os resultados encontrados por vários pesquisadores, descobriu-se que muitos modelos complexos foram desenvolvidos. Contudo, com o propósito de encontrar modelos preditivos mais precisos, não se pode negligenciar o fato de que modelos gestão de crédito pertencem a uma teoria aplicada. A redução na taxa de erro nas previsões é primordial para esses modelos, entretanto, esses resultados não têm significado ou valor se não forem aplicados na previsão de dificuldades financeiras de forma prática.

A complexidade de muitos modelos pode impedir que usuários finais - que nem sempre são especialistas em computação – resistam a sua utilização nas organizações. Esses modelos servem como uma ferramenta gerencial que podem interferir diretamente nas receitas de uma corporação, portanto, os gestores também poderão oferecer resistência na adoção de novas ferramentas nas operações corporativas, caso não compreendam claramente o mecanismo interno do modelo.

Khashman (2010); Wang et al. (2011), Wang; Ma (2011) argumentam que as pesquisas deveriam levar esse aspecto em consideração ao desenvolver modelos. Também, de acordo com esses autores, mais trabalhos deveriam se concentrar em projetar, treinar e implementar sistemas com mais resultados explicativos, como por exemplo o motivo pelo qual um pedido de crédito foi rejeitado.

Zhang et al. (2010); Kim; Kang (2010); Wang et al. (2011); Finlay (2011); Sun; Jia; Li (2011); Brown; Mues (2012); Yu et al (2016); Sun et al. (2017) sugerem que uma maneira de aprimorar a precisão de classificação seria utilizando um conjunto de técnicas complementares para uma tarefa específica de aprendizado. Ao contrário da abordagem mais frequente à modelagem, em que apenas um forte modelo preditivo é construído. Por essa razão, sugere-se que novos estudos sejam desenvolvidos utilizando métodos de *boosting* e *bagging*, que são baseados em uma estratégia sequencial onde vários outros modelos realizam variadas tarefas com um só propósito.

Um fato interessante confirmado em toda a pesquisa é a falta de estudos relevantes utilizando informações de mercados emergentes. Embora o portfólio final tenha selecionado dois estudos de países emergente considerados relevantes, mais pesquisas dessas regiões não foram capazes de produzir ou provar relevância científica. Portanto, a partir dos resultados aqui encontrados, sugere-se fortemente que mais estudos devam ser desenvolvidos

considerando informações de empresas desses mercados. Essa abordagem deve atrair atenção, pois, essas regiões são destino de investimentos de vários países e suas instituições. A América Latina, o Caribe e a África têm potencial para novas pesquisas.

Maioria dos estudos, nesta revisão, fez uso de mais de uma base de dados na construção de seus modelos. Dado esse fato, o uso de várias bases de dados com diferentes proporções de dados de treinamento, testes e a validação cruzada provavelmente contribuirão para uma melhor compreensão do desempenho dos modelos, oferecendo respaldo a conclusões mais confiáveis.

Outro aspecto observado nesta pesquisa foram as variáveis escolhidas nos estudos. O portfólio final mostrou que 48% dos artigos se basearam em informações de balanço patrimonial e 46% em dados de instituições financeiras. Lin, Liang e Chen (2011) construíram modelos usando fatores relacionados à governança corporativa e sugeriram considerar outras variáveis não-financeiras potencialmente influentes, tais como: participação de mercado, estilo de gestão e perspectiva da indústria, etc.

Kim (2011) também o sugere o uso de variáveis qualitativas como: capacidade de gestão, reputação, tipo de propriedade, investimentos futuros, etc., como uma opção de variáveis a serem considerada em estudos futuros. Ele sugere que devido à influência de condições macroeconômicas sobre o desempenho corporativo e do consumidor, espera-se que os modelos que incorporam variáveis macroeconômicas e não somente financeiras, melhorem a capacidade de previsão, especialmente quando a quantidade de informações sobre o perfil do cliente, seja ele pessoa física ou jurídica, está constantemente sendo gerada. Já Guo et al. (2016b) fez uso de variáveis não financeiras quando propõe o uso de informações de mídias sociais em seu construto.

Após a conclusão da primeira etapa desta pesquisa, deve-se destacar suas limitações. Entre elas, deve-se mencionar (1) seleção de bases de dados, pois o leitor poderia questionar quanto ao uso de um maior número de base indexadoras de artigos, (2) o tempo de investigação - estudos publicados após 1º de janeiro de 2008 até 31 de setembro de 2017, poderiam ter sido estendidos por um tempo maior, (3) críticas à escolha de palavras-chave que delimitaram o início da pesquisa, poderiam ser revistas ou complementadas, em uma lógica booleana mais complexa e (4) a formulação de outras questões e objetivos, permitindo assim a identificação de mais oportunidades de pesquisa

CAPÍTULO 3: USO DO MODELO *EXTREME GRADIENT BOOSTING* PARA PREVER DIFICULDADES FINANCEIRAS EM EMPRESAS LATINO-AMERICANAS

A previsão de dificuldades financeiras (FD) tem sido amplamente estudada em finanças corporativas por causa de seu impacto na própria sobrevivência e desenvolvimento da empresa bem como a decisão de investidores externos e credores (SUN; LI, 2009). Por sua vez, quando empresas são incapazes de pagar suas obrigações por dificuldades financeiras, aumentam as chances de bancos não receberem seus empréstimos e isso traria problemas para todo o sistema financeiro (SUN et al., 2017)

Sun et al. (2017) afirmam que em situações como essas é necessário o desenvolvimento de modelos eficientes de previsão de FD, os quais podem ajudar tanto as empresas a melhorar a gestão de riscos, como bancos a tomarem decisões de crédito de forma mais precisa e assertiva. O desenvolvimento de sistemas que são capazes de prever inadimplência e dificuldades financeiras são imperativos para que ambas as partes (credor e devedor) possam tomar ações tanto preventivas quanto corretivas. (WANG; WANG; LAI, 2005; LAI et al., 2006).

Em um amplo estudo Jones, Johnstone e Wilson (2015) compararam o desempenho de modelos tradicionais (Logit / Probit e Análise Discriminante) a modelos de aprendizagem de máquina, como Redes Neurais, *Support Vector Machine* e técnicas mais recentes como *generalized boosting*, *AdaBoost* e *Random Forest*. Em seu artigo, eles demonstraram que o último superou todos os outros métodos.

Além de sua eficiência na análise de dados, a aprendizagem de máquina possui aspectos que devem ser pesquisados com mais enfoque para fomentar sua adoção em empresas. Wang e Ma (2011) apontam que os modelos devem combinar precisão e usabilidade. Chen et al. (2011) sugeriu a melhora da interpretabilidade dos *ensambles* como direção de pesquisa importante, pela falta de conclusões satisfatórias. Bae (2012) recomendou explorar e construir modelos com diferentes bases de dados, posto que esta é uma questão delicada, uma vez que os bancos podem oferecer restrições em divulgar as informações de seus clientes.

Guo *et al.* (2016a) seguiu em outra direção insistindo que os modelos têm confiado apenas em variáveis numéricas e financeiras, dessa forma, recomendou-se experimentar variáveis não financeiras, quais sejam: fatores relacionados à governança corporativa (por exemplo, capacidade de gestão, reputação, tipo de propriedade, planos futuros, etc.),

condições macroeconômicas do desempenho corporativo e do consumidor e até mesmo dados de mídias sociais.

Kim e Kang (2010), Finlay (2011), Brown e Mues (2012), Tsai, Hsu e Yen (2014) e Kim, Kang e Kim (2015) recomendaram o desenvolvimento de modelos levando em consideração os métodos de *boosting* e *bagging*, que são baseados em uma estratégia construtiva de análise.

A partir das lacunas encontradas, este presente estudo decidiu por testar o modelo *Extreme Gradient* Boosting (Xgboost), um método de aprendizagem de máquina recente usado para problemas de aprendizado supervisionado (CHEN; GUESTRIN, 2016), onde o termo *Gradient Boosting* foi proposto por Friedman (2001). Xgboost é um aprimoramento e baseado no modelo de Friedman (2001). A escolha desse modelo repousa sobre eficiência demonstrada, precisão e praticidade de seu algoritmo. Além disso, sua capacidade de realizar diversos cálculos em bases com um extenso volume de dados, mesmo em um computador comum, faz com que a escolha desse modelo seja ainda mais promissora. O algoritmo é construído de tal forma que ele possui, também, recursos adicionais para realizar validação cruzada e consiga exibir as variáveis mais impactantes. Além da análise dos dados, alguns outros esforços interessantes foram inseridos no algoritmo Xgboost, tais com otimização de memória, otimização de cache e melhoria em termos de modelo em si, com programação de memória externa, abstrações distribuídas, o qual ajuda a entender qual algoritmo é adequado para qual caminho.

Neste estudo, confirmou-se seu desempenho e precisão quando comparados à modelos mais usados (Regressão Logística e *Random Forest*). O Xgboost obteve uma taxa de acerto de 96,05% contra 95,10% do *Random Forest* e 65,12% da Regressão Logística.

Esses achados devem contribuir para a literatura sobre predição do risco de crédito de algumas maneiras. Isso se torna um fato importante pois os códigos do algoritmo estão disponíveis ao público para que os bancos e futuros investidores, analistas e gestores possam evitar o pré conceito “caixa preta” que os modelos complexos têm. Finalmente, esta pesquisa tem a intenção de incentivar a aliança entre finanças e a ciência da computação na busca de um processo estruturado e preciso na tomada de decisões.

3.1 Revisão da Literatura

O Acordo de Basileia II determina que empresas divulguem práticas de gerenciamento de risco, o que exige modelos mais confiáveis e precisos para classificar e quantificar essas probabilidades. (BIS, 2006). Por esse motivo a adoção de algoritmos de Aprendizagem de

Máquina, subcampo das ciências da computação que se refere ao estudo da teoria de reconhecimento de padrões e aprendizagem computacional utilizando a inteligência artificial.

Estudos compararam o desempenho de diferentes métodos. Alfaro et al. (2008) confirmaram que *AdaBoost* supera Redes Neurais e que seu teste de erro foi de 8,898% contra 12,712% das redes neurais. Heo e Yang (2014) compararam várias taxas de sucesso do algoritmo de aprendizagem de máquina e testaram-no contra o famoso Z-score de Altman: AdaBoost (78,5%), ANN: (77,1%) SVM (73,3%), DT (73,1%) e Altman Z -escore (51,3%).

Na Tabela 4, pode-se perceber que vários autores se esforçaram com diferentes algoritmos. Esses pesquisadores destacaram a capacidade dos modelos, mas também apontaram suas desvantagens, como sua natureza obscura, maior carga computacional, propensão ao *sobreajuste* e natureza empírica da construção.

Tabela 4 - Métodos de Aprendizagem de Máquina utilizados em risco de crédito.

Autores	**Algoritmos adotados**
Tsai and Wu, 2008; Chauhan, Ravi and Chandra, 2009; Kim and Kang, 2010; Du Jardin, 2010; Chuang and Huang, 2011; Marcano-Cedeño et al., 2011; Jeong, Min and Kim, 2012; Blanco et al., 2013; Lee and Wu Sung, 2013; López and Sanz, 2015; Zhao et al., 2015; Yu, Yang and Tang, 2016	Redes Neurais
Sun and Li, 2012; Wang and Ma, 2012; Hens and Tiwari, 2012; Hsieh et al., 2012; Harris, 2015; Danenas and Garsva, 2015; Sun et al., 2017	Support Vector Machine
Li, Sun and Wu, 2010; Cho, Hong and Ha, 2010; Zhang et al., 2010; Gepp, Kumar and Bhattacharya, 2010; Wang et al. 2012; Kim and Upneja, 2014	Decision Trees
Sun, Jia and Li, 2011; Wang and Ma,2011; Wang, Ma and Yang, 2014; Kim and Upneja, 2014; Heo and Yang, 2014; Kim, Kang and Kim, 2015; Sun et al., 2017	Boosting Algorithms

Fonte: Resultados da pesquisa.

Embora quase todos os métodos possam ser usados para avaliar o risco de crédito, recentemente - devido à crescente complexidade e tamanho das bases de dados - pesquisadores têm experimentado diferentes classificadores e técnicas, que integram dois ou mais métodos de classificação. Essas abordagens têm mostrado maior precisão na previsibilidade do que métodos individuais e tem atraído bastante atenção na avaliação de risco de crédito. Alguns destes exemplos são a Técnica Discriminante Neural (Lee *et al.*, 2002), *neuro-fuzzy* (PIRAMUTHU, 1999; MALHOTRA; MALHOTRA, 2002) e fuzzy-SVM (WANG et al., 2005).

Em um banco de dados com informações financeiras de clientes, têm-se um cenário em que o número de observações associadas a uma classe é bem maior do que aquelas pertencentes à outra classe. Por essa razão, há uma questão de desproporção de informações, que Brown e Mues (2012) examinaram em seu estudo investigando vários tipos de algoritmos, cujos resultados demonstraram que *Random Forest* - um algoritmo baseado em árvores de decisão - e *Gradient Boosting* desempenharam relativamente bem em conjuntos de dados desbalanceados.

3.2 O modelo

Dado as vantagens e desvantagens de diversos modelos, este estudo buscou testar um modelo de classificação que pudesse determinar a probabilidade de inadimplência dos clientes de cartão de crédito, além de monitorar quais variáveis devem ser observadas para antecipar o evento. Wang e Ma (2011) aplicaram o *RS-Boosting* e obtiveram melhores resultados principalmente na redução do erro do tipo II, mas a partir do seu trabalho detectou-se que a interpretabilidade dos resultados é outra direção de pesquisa importante ainda pouco explorada.

O modelo Xgboost mostrou evidências de sua precisão em mais da metade das soluções vencedoras em desafios de aprendizagem de máquina sediados pela Kaggle, uma comunidade on-line de cientistas de dados e aprendizes de máquinas, de propriedade da Google Inc. (He, 2016). Nesse algoritmo pode-se verificar que as conclusões de Brown e Mues (2012) corroboram com esta pesquisa e que este estudo não seria capaz de esgotar as possibilidades de uso modelo.

Em finanças, a aplicação desse modelo é relativamente nova. He, Zhang e Zhang (2018) compararam o desempenho do Xgboost com outros modelos de escore de crédito e obtiveram os melhores resultados de classificação em quatro dos seis bancos de dados utilizados, o que indica que ele tem excelente desempenho. Carmona, Climent e Momparler

(2018) testaram o modelo para prever falências no setor bancário dos EUA e concluíram que o XgbGBoost tem maior poder preditivo do que os métodos Regressão Logística e *Random Forest*. Xia et al. (2018) apontam a superioridade do modelo como meta-classificador. Xia et al. (2017) destacaram comparações com diferentes modelos de base e mostraram a superioridade do modelo baseados no Xgboost em termos de desempenho preditivo.

Contrapondo algoritmos que utilizam o método *bagging*, o qual constrói modelos de forma paralela, a abordagem do *boosting* é construir modelos de forma sequencial. O Xgboost usa *K* classificadores f_k (x) para aproximar do modelo final F_k (x) e minimizar a função de custo fornecida. O método de *gradient descent* calcula a derivada parcial e tenta otimizar a função de custo buscando a mínima (local) ajustando diferentes valores de coeficientes para minimizar o erro de forma iterativa. Esta função de custo mede quão bem o modelo se ajusta aos dados atuais e o processo de *boosting* continua até que a redução da função de custo se torne limitada. (CHEN; GUESTRIN, 2016)

O objetivo final de sua aprendizagem é determinar um roteiro, onde $\boldsymbol{y}$ é a previsão esperada e $\boldsymbol{x}$ são os vetores de características. (XIA et al., 2017). Para construir tal mapa a ser seguido, o modelo proposto requer múltiplos parâmetros que devem ser pré-definidos e o controle da combinação adequada de parâmetros é fundamental para otimizar e melhorar o modelo (CHEN; BENESTY, 2016). Parâmetros gerais são discutidos nos seguintes tópicos:

3.3.1 Parâmetros

• *Número de rodadas ou número máximo de iterações*: o número ideal de rodadas ou árvores necessárias no modelo;

• *Profundidade ou tamanho máximo de uma árvore:* é o número de divisões em cada árvore. Ele é usado para controlar o *sobreajuste*, pois uma profundidade maior permite que o modelo aprenda com padrões que são característicos de uma amostra específica;

• *Taxa de aprendizagem*: introduzida inicialmente por Friedman (2002), consiste em um número positivo (variando de 0 a 1) que determina a rapidez com que o algoritmo se adapta ou a contribuição de cada árvore para o modelo. Um valor baixo significa que o modelo é mais robusto ao *sobreajuste*.

• *Gamma*: redução mínima de perda necessária para fazer a próxima divisão em um nó de cada árvore. Quanto maior o seu valor, mais conservador será o algoritmo;

• *Coluna e amostra de observação*: uma proporção da sub amostra de variáveis e observações ao construir cada árvore. A coluna e amostra de observação denota a fração de

variáveis e observações, que devem ser amostradas aleatoriamente para cada árvore. Seu valor varia de 0 a 1, evita *sobreajuste* e acelera os cálculos do algoritmo.

• *Minimum child weight*: indica a soma mínima do peso da instância necessária em um nó. Se a etapa de separação de árvores resultar em um nó folha cuja soma de peso precedente é menor que o valor atribuído a esse parâmetro, o processo de construção interromperá subdivisões adicionais.

• *Termo de regularização ou penalidade nos pesos:* o termo de regularização controla a complexidade do modelo que ajudar a evitar o sobre ajuste.

3.3 Modelos de referência

O objetivo principal não é apenas contrapor diferentes métodos de aprendizagem de máquina, mas também evidenciar um modelo recentemente desenvolvido e, assim, motivar pesquisas unindo áreas de finanças e ciência da computação. O modelo Xgboost foi comparado com um método estatístico convencional (regressão logística) e um conhecido algoritmo de aprendizagem de máquina (Random Forest). Além de seu desempenho, o Xgboost possui outro aspecto importante que é a identificação e exibição de variáveis mais importantes para o modelo.

A regressão logística é uma das técnicas mais aceitas e utilizadas em termos teóricos, uma vez que duas classes distintas (boas ou ruins) foram definidas de antemão. (KIM, 2011; LI; SUN, 2011b; Li *et al.,* 2011). Dado um conjunto de treinamento de *N* registros $D = \{(\mathrm{xi}, \mathrm{yi})\}_{\mathrm{i}=1}^{N}$, com variáveis independentes $\mathrm{x}_i \in R^N$ e variáveis binárias dependentes correspondentes $\mathrm{y}_i \in \{0,1\}$, a regressão logística busca classificar (*LOG*) e estimar a probabilidade P (y = 1 | x) de clientes adimplentes e inadimplentes da seguinte maneira:

$$\boldsymbol{P(y = 0|\, x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}} \qquad \textbf{(1)}$$

A Árvore Aleatória é um método baseado em árvores binárias selecionada aleatoriamente que emprega *bagging* para agrupar e obter subconjuntos diversificados de todo o conjunto de dados de treinamento e construir árvores individuais. O algoritmo é um uma técnica de classificação que consiste em uma seleção de classificadores estruturados em árvores $\{h(\mathbf{x}, \Theta_k), k = 1,...\}$ onde $\{\Theta_k\}$ são vetores aleatórios distribuídos independentes e cada árvore lança um voto unitário para as demais classes de variáveis $\boldsymbol{x}$. Esse algoritmo adiciona uma aleatoriedade extra ao modelo, pois, ao invés de procurar pela melhor característica dos dados ao fazer a partição dos nodos, ele busca a melhor característica em um subconjunto aleatório das características. (BREIMAN, 2001).

3.4 Configuração do experimento

3.4.1 Base de dados

A seleção da base de dados deu-se após a constatação de que estudos com informações de mercados emergentes são poucos e possuem pouca relevância científica. A Figura 7 no capítulo 2 atesta essa afirmação. A figura mostra de forma gráfica os artigos selecionados para o portfólio final. A seleção das variáveis independentes foi feita em consonância com estudos desta seleção.

A amostra da pesquisa é composta por todas as empresas Latino Americanas (ativas e canceladas) pertencentes à base de dados Economática, no período de 2000 a 2017. As empresas canceladas foram incluídas como forma de mitigar o viés de sobrevivência, entretanto, empresas do setor financeiro foram excluídas da amostra devido às características contábeis e setoriais específicas de tal segmento.

A variável de interesse foi operacionalizada como uma medida representativa da gestão de endividamento das empresas (PINDADO; RODRIGUES; DE LA TORRE, 2008). Assim, este estudo utiliza um critério financeiro na determinação de uma variável, pois definições de dificuldades financeiras baseadas na falha da empresa em cumprir suas obrigações financeiras é consistente com uma abordagem *ex-ante,* ou seja, adota uma definição de dificuldade financeira que avalia a capacidade da empresa para satisfazer suas obrigações financeiras (SANZ; AYCA, 2006).

Assim, o presente estudo classifica uma empresa como com dificuldades financeiras, não só quando essa decreta falência, mas também quando ambas as seguintes condições são satisfeitas: 1) lucro antes de juros e impostos depreciação e amortização (EBITDA) são inferiores às suas despesas financeiras para dois anos consecutivos, levando a empresa a uma situação em que não pode gerar fundos suficientes de suas atividades operacionais para cumprir com suas obrigações financeiras; (2) uma queda em seu valor de mercado ocorre entre dois períodos consecutivos. (MANZANEQUE; MERINO; PRIEGO, 2016). Assim sendo, esta pesquisa considera uma empresa com dificuldades financeiras no ano que segue imediatamente a ocorrência desses dois eventos, sendo que este critério divide a amostra em dois grupos. Essa condição é representada por uma variável dependente binária que se tem valor “1” (um) para empresas com dificuldades financeiras e “0” (zero) para aquelas que não apresentassem tal dificuldade. O Quadro 1 apresenta a definição das condições necessárias para determinação da variável dependente em $t+1$, demonstrando, inclusive, exemplos de estudos que utilizaram as mesmas medidas.

Quadro 1 - Condicionantes para variável dependente.

Variável	Descrição	Cálculo	Autores	Fonte
EBITDA< Desp.Fin	EBITDA menor que Despesas Financeiras por dois anos consecutivos.	EBITDA < Desp.Fin.	Pindado et al. (2008)	Economática
FALLMKT.VAL	Queda no valor de mercado entre dois períodos consecutivos.	$MKT.Value_{t-1} < MKT.Value_t$	Pindado et al. (2008)	Economática

Fonte: Elaborado pelo autor.

A seleção das variáveis independentes também foi feita em consonância com estudos mais relevantes identificados no portfólio de artigos produzidos pelo processo de revisão sistemática da literatura, o qual é descrito no capítulo 2. O Quadro 2 resume as informações pertinentes as variáveis independentes do estudo.

Quadro 2 – Definição das variáveis independentes do estudo.

Variável[a]	Descrição[b]
CATA	Ativo Circulante/ Ativo Total
COMENDIV	Composição do Endividamento: Passivo Circulante/ (Passivo Circulante+Exigivel a Longo prazo)
ENDV	Passivo Circulante + Passivo Não Circulante/Ativo total
FATA	Ativo Permanente/ Ativo Total
GR/Sales	Receita Bruta/ Vendas
GRWTA	Taxa de Crescimento ativo total
LIQCOR	Ativo Circulante / Passivo Circulante
LIQ.IMED	Disponível / Passivo Circulante
LPA	Lucro Por Ação
MARGL	Margem Liquida (%)
Net.P/CAS	Lucro Liquido/ Ativo Total Circulante
Net.P/EQTY	Lucro Liquido/ Patrimonio liquido

Net.P/FIXA	Lucro Liquido/ Ativo Permanente
Net.P/TOTA	Lucro Liquido/ Ativo Total
PAS/PATL	Passivo/ Patrimonio Liquido
PATL/ATNCIR	Patrimônio Líquido/ Ativo não Circulante
PMREC	Prazo Médio de Recebimento (dias)
(a) Legenda para autores que também utilizaram a medida: Li Sun e Sun (2008), Li e Sun (2009), Li, Sun e Wu (2010), Li e Sun (2010), Li et al. (2010), Li et al. (2011), Lee et al. (2011) Li e Sun (2011), Jia et al. (2011), Li e Sun (2012), Chang et al. (2016), Wang et al. (2017), Fujita et al. (2017); **(b): Todos os dados foram obtidos pela base Economática.**	

Fonte: Elaborado pelo autor.

3.5 Definição da amostra

A análise contemplou empresas ativas e canceladas pertencentes à base de dados Economática, no período de 2000 a 2017. A não exclusão de empresas canceladas foi com o intuito de evitar o viés de sobrevivência, isto é, este viés seria maior caso se considerasse somente empresas ativas ao final de 2017 (IQUIAPAZA; LAMOUNIER; AMARAL, 2008). As empresas do setor financeiro foram excluídas da amostra, devido às características contábeis e setoriais específicas de tal segmento, conforme (CAMPOS; NAKAMURA, 2015). As empresas que não possuíam qualquer informação também foram eliminadas da amostra. A Tabela 5 demonstra o total de empresas na amostra final.

Tabela 5 - Definição da amostra do estudo

Classificação	Quantidade de Empresas
Empresas Ativas	702
(+) Empresas Canceladas	353
(=) Amostra Final	1055

Fonte: Elaborada pelo autor.

3.5.1 *SMOTE: Synthetic Minority Over-sampling Technique*

As performances dos algoritmos de aprendizado de máquina são tipicamente avaliadas usando métricas de precisão. No entanto, estas podem ser inapropriado quando os dados estão desbalanceados (CHAWLA et al., 2002). A literatura acerca do tema aborda a questão da

desproporção de classes de duas maneiras. Uma é atribuir custos distintos à amostra de treinamento (PAZZANI *et al,* 1994; DOMINGOS, 1999). A outra é produzir uma amostra a partir do conjunto de dados original, seja por *oversampling* ou *undersampling* da classe majoritária (JAPKOWICZ, 2000).

Na amostra final de empresas analisadas ao longo de todo período, os resultados mostram que 92% das observações são de situações onde a empresa não foi classificada com dificuldades financeiras no ano seguinte e somente 8% das observações identificadas com tal descrição. Desse modo, a amostra final foi balanceada utilizando o método proposto por Chawla et al. (2002), no qual novas observações sintéticas são criadas baseadas nas observações minoritárias existentes. O balanceamento dos dados pelo SMOTE foi alcançado utilizando a plataforma *Microsoft Azure Machine Learning Studio* , resultando em uma base balanceada com 52% de observações classificadas como sem dificuldades financeiras e 48% com dificuldades financeiras.

3.6.1 Seleção de variáveis

Gujarati e Porter (2011, p. 39) destacam que na análise de dados deve-se entender a dependência estatística entre as variáveis, evitando as relações não funcionais ou determinísticas. O algoritmo Xgboost é um algoritmo baseado em árvores de decisão que utiliza o método gradiente descendente para construí-las otimizando seus pesos. O núcleo do algoritmo é otimizar o valor da função objetivo. (CHEN; GUESTRIN, 2016)

Ao contrário do uso de coeficientes para calcular a capacidade de previsão que cada variável possui, o método gradiente descendente constrói árvores sequenciais para obter de forma eficaz os escores, indicando a importância de cada característica para o modelo de treinamento. Quanto mais uma variável for utilizada na criação das árvores, maior será seu peso. O algoritmo considera a importância por “ganho”, “frequência” e “cobertura” (FRIEDMAN; HASTIE; TIBSHIRANI, 2001). O ganho é o principal fator de referência para importância de uma variável nos galhos das árvores formadas.

Dessa maneira, foram utilizadas 18 variáveis independentes disponíveis na base Economática conforme tabela 6. Essas variáveis foram selecionadas a partir do trabalho dos autores de maior relevância para o portfólio final de artigos elaborado pela revisão da literatura no capítulo anterior. A equação 2, mostra algoritmo com todas variáveis que foram inclusas no modelo. No Apêndice II, no final dessa pesquisa, foi disponibilizado o script do modelo Xgboost para o software R, bem como suas extensões necessárias.

build_model(model_func = xgboost_binary, formula = FINDISTR ~ PMREC + ENDV + LIQIMED + CATA + COMENDIV + LIQCOR + LPA + PASPATL + PATLATÑCIR + FATA + GRSales + GRWTA + NetPCAS + NetPEQTY + NetPFIXA + NetPTOTA + MARGL + MARKVAL)

(2)

3.6 Resultados encontrados

A Tabela 6 demonstra as estatísticas descritivas do estudo, evidenciando as informações pertinentes ao número de observações, média, desvio-padrão e valores mínimos e máximos. É possível observar que o LPA médio no período estudado é negativo, neste sentido, pode-se inferir que empresas estavam operando com margens baixas, consequentemente, acumulando prejuízos, de alguma forma. De forma complementar, nota-se que o índice COMENDIV é de, em média, 72,78% e revela que grande parte da dívida média total com terceiros é exigível no curto prazo.

Tabela 6 - Estatística descritiva dos dados.

Variável	Observações	Média	Desvio Padrão	Mínimo	Máximo
PMREC	17.282	229,2898	2998,601	-8423,438	326606
ENDV	17.282	1,114207	7,401676	0	487,224
LIQIMED	17.282	0,828647	3,319415	-0,9997084	159,2569
CATA	17.282	0,3727032	0,2037497	0	1
COMENDIV	17.282	0,7278591	0,2740701	0	1
LIQCOR	17.282	1,831159	3,344719	0	160,2569
LPA	17.282	-5,539112	84,37737	-2364,538	2661,399
PASPATL	17.282	1,447159	24,59009	-901,5452	1148,586
PATLATÑCIR	17.282	1,30E+08	8,90E+09	-753,173	8,60E+11
FATA	17.282	0,5481444	0,2752177	0	0,9999502
GRSALES	17.282	5377,712	706899,7	-3,948443	9,29E+07
GRWTA	17.282	7,204386	848,9291	-0,9643702	111083,2
NETPCAS	17.282	-1,331114	29,30583	-1207,017	618,7882
NETPEQTY	17.282	-0,0036559	8,458128	-1066,615	109
NETPFIXA	17.282	1,87E+07	1,61E+09	-2,56E+08	1,88E+11
NETPTOTA	17.282	-0,0465863	0,6887682	-46,33359	11,88856
MARGL	17.282	-1445,015	33932,04	-1899698	132037,5
MARKVAL	17.282	1,65E+09	7,11E+09	0	2,43E+11

CATA - Ativo Circulante/ Ativo Total; COM.ENDIV - Composição do Endividamento: Passivo Circulante/ (Passivo Circulante+Exigivel a Longo prazo); EBIT/Desp.Fin - EBIT/ Despesas Financeiras; ENDV - Endividamento: Passivo Circulante + Passivo Não Circulante/Ativo total; FATA - Ativo Permanente/ Ativo Total; GR/Sales - Receita Bruta/ Vendas; GRW.TA - Taxa de Crescimento ativo total; LIQ.COR - Ativo Circulante / Passivo Circulante; LIQ.IMED - Disponível / Passivo Circulante; LPA - Lucro Por Ação; MARG.L - Margem Liquida (%); MARK.VAL - Valor de Mercado; Net.P/C.AS - Lucro Líquido/ Ativo Total Circulante; Net.P/EQTY - Lucro Líquido/ Patrimônio Líquido; Net.P/FIX.A - Lucro Líquido/ Ativo Permanente; Net.P/TOT.A - Lucro Líquido/ Ativo Total; PAS/PAT.L - Passivo/ Patrimônio Líquido; PAT.L/AT.ÑCIR - Patrimônio Líquido/ Ativo não Circulante; PMREC - Prazo Médio de Recebimento (dias)

Fonte: Resultados da pesquisa.

O Apêndice III demonstra a matriz de correlação das variáveis do estudo. É possível verificar uma associação positiva entre a variável dependente e o FATA. Essa variável mostra a relação de ativos fixos em relação ao ativo total. Em outras palavras, o quanto de capital é investido em ativos que ajudam a aumentar a receita.

Os parâmetros do modelo Xgboost foram calibrados para chegar ao melhor modelo e identificar uma estrutura nos dados. Embora o melhor fracionamento dos dados e ajuste dos parâmetros demande futuras pesquisas, este tópico em específico está além do escopo deste estudo.

Concluída a descrição dos dados, cinco métricas de avaliação tradicionais - precisão, erro do tipo I, erro do tipo II, área sob a curva ROC (AUC) e teste de Kolmogorov-Smirnov (KS) - foram empregadas para avaliar o desempenho dos modelos. A principal métrica de avaliação será AUC (Área sob a curva), que é uma medida alternativa de capacidade de discriminação baseada na curva ROC. A curva ROC mostra valores da taxa de verdadeiros positivos (TPR) contra os valores da taxa de falsos positivos (FPR) em várias configurações de limite. A taxa de verdadeiro positivo também é conhecida como sensibilidade e a taxa de falso positivo é conhecida como especificidade e pode ser calculada da seguinte maneira (1 - especificidade). Em outras palavras, o escore AUC mede quão bem o modelo discrimina entre as duas classes.

Os parâmetros do modelo Xgboost foram calibrados para chegar ao melhor modelo e identificar uma estrutura nos dados. Embora o melhor fracionamento dos dados e ajuste dos parâmetros demande futuras pesquisas, este tópico em específico está além do escopo deste estudo.

Concluída a descrição dos dados, cinco métricas de avaliação tradicionais - precisão, erro do tipo I, erro do tipo II, área sob a curva ROC (AUC) e teste de Kolmogorov-Smirnov

(KS) - foram empregadas para avaliar o desempenho dos modelos. A principal métrica de avaliação será AUC (Área sob a curva), que é uma medida alternativa de capacidade de discriminação baseada na curva ROC. A curva ROC mostra valores da taxa de verdadeiros positivos (TPR) contra os valores da taxa de falsos positivos (FPR) em várias configurações de limite. A taxa de verdadeiro positivo também é conhecida como sensibilidade e a taxa de falso positivo é conhecida como especificidade e pode ser calculada da seguinte maneira (1 - especificidade). Em outras palavras, o escore AUC mede quão bem o modelo discrimina entre as duas classes.

Recobrando a revisão da literatura, não foi possível identificar estudos relevantes utilizando o modelo proposto até a data limite de 2017. Dado esse fato, os parâmetros foram ajustados de acordo com Carmona, Climent e Momparler (2018) conforme descrito na tabela 8. A intenção de calibrar e controlar parâmetros é a de evitar o sobre ajuste do modelo aos dados e garantir sua generalização. O modelo foi treinado com 1000 iterações ou rodadas, profundidade máxima de árvore de 5, taxa de aprendizado de 0,1, gama de 0, proporção de subamostras de 0,8, *minimum child weight* de 1 e, por último, valor de regularização de 0.

Tabela 7 - Parâmetros do XGboost

Parâmetros	Valores
Number of Iterations	1000
Maximum depth	5
Learning rate	0,1
Gamma	0
Observation sample	0,8
Min. Child Weight	1
Regularization	0

Fonte: Carmona, Climent e Momparler (2018).

Depois de avaliado o modelo, ele foi testado e, na tabela 10, os resultados das métricas de desempenho são apresentados. O desempenho do modelo foi avaliado em um conjunto de dados diferente do utilizado para treiná-lo. Assim, aleatoriamente as observações foram divididas em 70% para treinamento e 30% para teste do modelo. No fragmento maior de dados, ele foi treinado e ajustado, enquanto o fragmento menor foi utilizado para testá-lo. Também, para os modelos concorrentes, Regressão Logística e Árvore Aleatória, a partição dos dados seguiu o mesmo raciocínio, 70% para treinamento e 30% para teste dos modelos, seguindo (LI et al., 2017; XIA; YANG; ZHANG, 2018).

No tocante às métricas de avaliação, quatro medidas diferentes são expostas para todos os três modelos. A matriz de confusão na tabela 8 traz quatro elementos básicos: verdadeiros positivos (TP) indicando que a previsão de adimplentes é consistente com seu valor real; falsos negativos (FN) significa que o resultado da previsão da amostra é classificado como inadimplentes, mas seu rótulo real indica adimplente. Da mesma forma, falsos positivos (FP) são aquelas amostras de inadimplentes classificadas como adimplentes e aquelas dentro de amostras de inadimplentes corretamente previstas como inadimplentes são rotuladas como negativas verdadeiras (TN).

Tabela 8 - Matriz de confusão.

		Previsto	
		Positivo	Negativo
Real	Positivo	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Negativo	Falso Positivo (FP)	Verdadeiro Negativo (FN)

Fonte: Elaboração própria.

O *ACC, dada pela equação 3,* é a medida de precisão do modelo em comparação aos dados gerais. É a razão entre as unidades corretamente classificadas e o número total de previsões feitas pelos classificadores. O *ACC* é calculado da seguinte maneira:

$$\boldsymbol{ACC = \frac{TP+TN}{TP+FP+FN+TN}} \quad \textbf{(3)}$$

O *Under the ROC Curve* (AUC) mede a capacidade de um modelo binário, em outras palavras, uma maneira de representar a curva ROC em único valor, agregando todos os limiares da ROC calculando a "área sob a curva". O limite para duas classes é 0,5 e quanto mais próximo do valor 1, melhor o algoritmo é capaz de distinguir entre as classes.

O teste de Kolmogorov-Smirnov (KS) foi outra métrica usada para medir a adequação do modelo, usada para testar a normalidade, onde TPR significa taxa de positivos verdadeiros e FRP é taxa de falsos positivos. *KS* é um dos métodos não paramétricos mais úteis e difundidos para a comparação de duas amostras (ZHANG; PRIESTLEY; NI, 2018)

$$\boldsymbol{KS = \max_{t}(|TPR(t) - FRP(t)|)} \quad \textbf{(4)}$$

Por fim, as taxas de erro tipo I e II foram usadas como indicadores para explorar ainda mais a capacidade de o modelo classificar clientes em adimplente ou inadimplentes,

respectivamente. Neste experimento, a taxa de erro Tipo I denota a proporção de adimplentes classificados incorretamente, e a taxa de erro Tipo II refere-se à proporção de valores erroneamente classificados. Seus valores são calculados conforme as equações (4) e (5).

$$\boldsymbol{Erro\ I} = \frac{\boldsymbol{FP}}{\boldsymbol{FP+TN}} \quad \textbf{(5)}$$

$$\boldsymbol{Erro\ II} = \frac{\boldsymbol{FN}}{\boldsymbol{TP+FN}} \quad \textbf{(6)}$$

Para a base de dados deste estudo, o modelo Xgboost alcançou o melhor escore de *ACC* de 0,4763, com erro tipo I totalizando 0,0471 e 0,0312 de erro tipo II. A Árvore Aleatória tem desempenho relativamente próximo ao modelo proposto, com pontuação ACC de 0,4934 e 0,0411 para erro do Tipo I e 0,0,0571 para erro do Tipo II. O modelo Xgboost é particularmente útil para resolver problemas de classificação, pois minimiza os erros gerais ao gerar modelos baseados em erros em iterações de árvores anteriores. Quanto ao valor da estatística *KS,* o modelo Xgboost apresentou um resultado pouco melhor, 39,64% comprovando assim, sua promissora capacidade de descriminação entre duas classes de clientes.

SUN et al. (2017) utilizou em seu estudo o modelo *AdaBoost support vector machine,* um algoritmo que se baseia em erros de classificações anteriores para fazer a próxima classificação de forma também de forma sequencial, como o Xgboost. Em seu estudo, os autores aplicaram o modelo para prever situações de dificuldades financeiras de empresas chinesas e alcançaram uma acurácia de 93,88%. Em um estudo similar, Kim, Kang e Kim (2015) utilizam GMboost (Geometric Mean based Boosting) para prever falência em empresas coreanas e o modelo chega a uma acurácia de 82%.

Carmona, Climent e Momparler (2018) testaram os mesmos modelos propostos nesta pesquisa, contudo aplicado ao setor bancário e chegaram a resultados próximos dos encontrados por esta pesquisa. A tabela 9 mostra os valores alcançados pelos autores.

Tabela 9 - Resultados encontrados por Carmona, Climent e Momparler (2018)

Modelo	**Performance**	
	AUC	**Accuracy**
XGBoost	0.98	94.74%
Random Forest	0.93	92.11%
Regressão Logística	0.84	84.21%

Fonte: Carmona, Climent e Momparler (2018).

A tabela 10 traz as medidas de validação dos modelos testados nesta pesquisa e por meio dos resultados, pode-se perceber a capacidade e acurácia do modelo Xgboost, mesmo quando comparado a outros modelos e estudos.

Tabela 10 – Medidas de validação dos modelos comparados.

Modelo	***ACC***	***AUC***	***KS***	***Erro Tipo I***	***Erro Tipo II***
Reg. Logistica	0,6512	0,6751	0,1439	0,3108	0,3782
Árvores Aleatórias	0,9510	0,9502	0,3905	0,0411	0,0571
Xgboost	0,9605	0,9636	0,3964	0,0471	0,0312

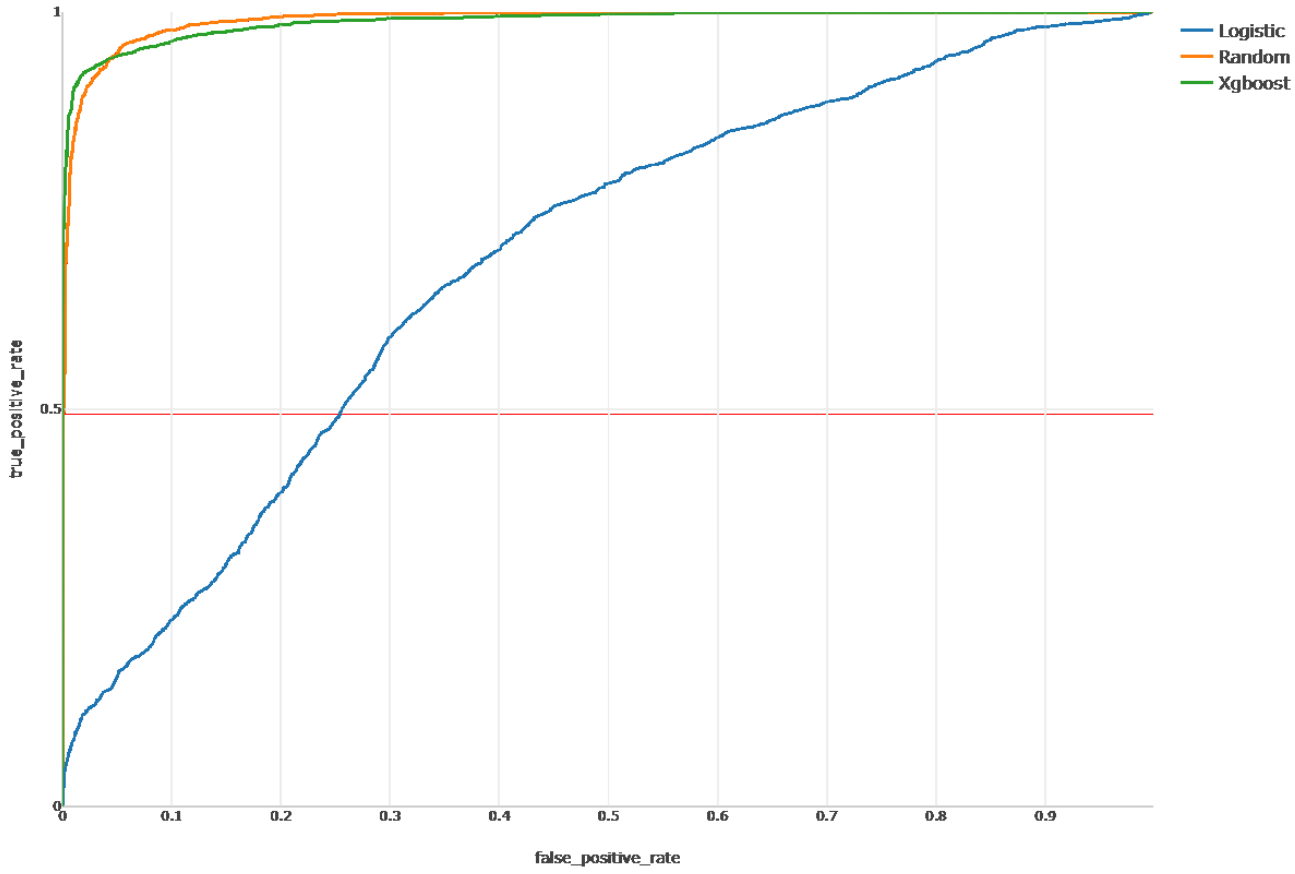
ACC=accuracy, AUC=Area Under the Curve, KS = Kolmogorov-Smirnov

Fonte: Resultados da pesquisa.

O erro tipo I está relacionado a não previsão de uma empresa com dificuldades financeira e o tipo II está relacionado a uma empresa que não apresenta esse quadro, todavia é rotulada como tal. Do ponto de vista financeiro, os riscos associados ao erro tipo I são mais onerosos do que Tipo II (WEST, 2000). Em outras palavras, nesta pesquisa pode ser visto como uma "perda de oportunidade de investimento versus a oportunidade de investimento" nas empresas analisadas.

A área sob a curva ROC (AUC) é um indicador importante, pois nos fornece uma medida da precisão total independente. O valor da área abaixo da diagonal (0,5 ou 50%) não tem validade, uma vez que são considerados como aleatórios. Entretanto, um valor que se aproxima do valor 1,0 ou 100% mostra o quão capaz é o modelo em fazer previsões corretas. Ela também, mostra de forma ilustrativa a capacidade de discernimento entre as classes. Na figura 13, os três modelos foram comparados. Pode-se dizer que o modelo Xgboost teve uma excelente performance, todavia, quando comparado ao modelo *Random Forest* a diferença é pequena, porém, inferior ao modelo proposto.

Figura 13 - Curva ROC para cada modelo testado.



Fonte: Resultados da pesquisa.

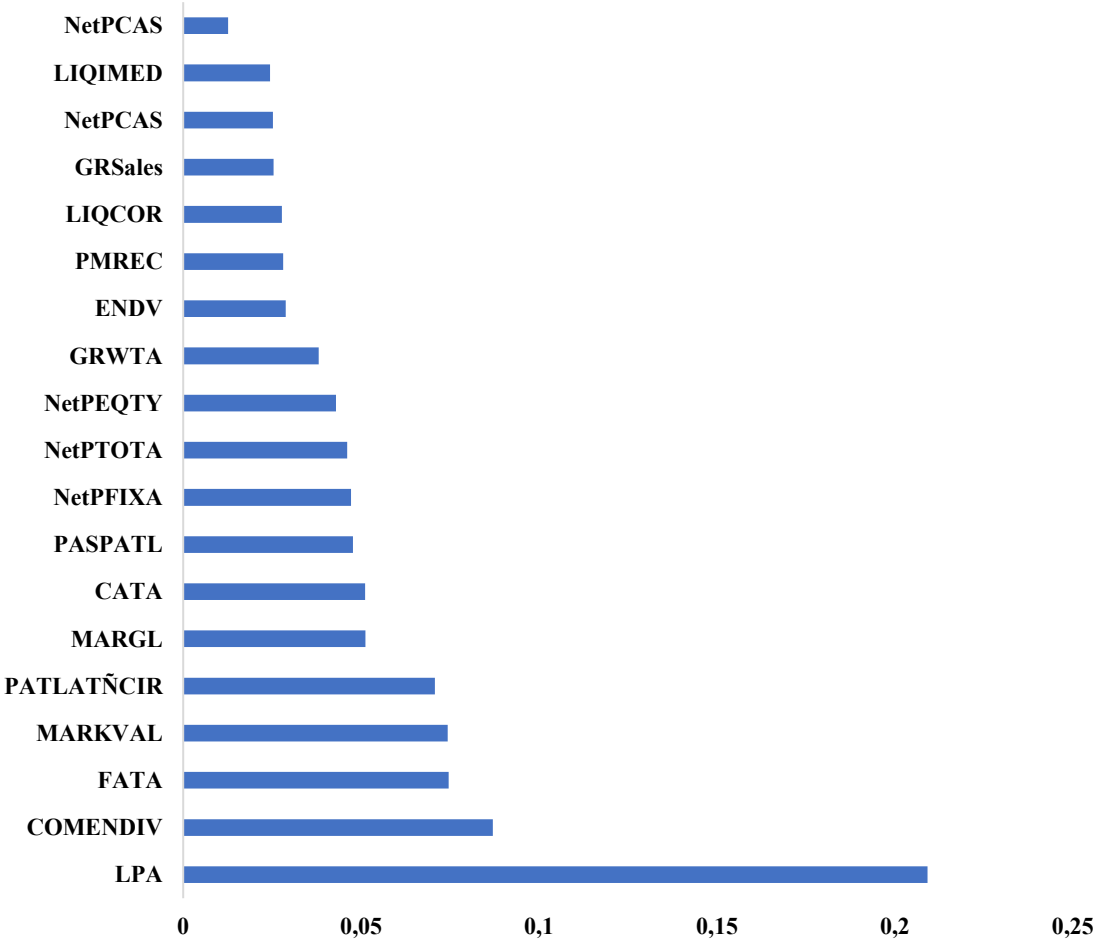
Outro aspecto importante do Xgboost é que seus resultados expõem as variáveis com maior influência sobre a previsão da variável dependente, corroborando assim, com ABDOU (2009). Em seu estudo o pesquisador sugere futuras investigações sobre quais variáveis podem indicar possíveis dificuldades financeiras com antecedência. Isso se torna uma questão importante porque empresas podem rastrear e monitorar características específicas de suas operações com intuito de evitar transtornos financeiros.

Na Figura 16 pode-se observar que as variáveis “LPA” e “COMENDIV” fornecem os maiores valores discriminatórios. A partir desse resultado, é possível inferir um alerta de que empresas já com dificuldades financeiras estariam comprometidas em reter recursos para o pagamento de juros e amortização das dívidas, o que desencadearia a correlação entre endividamento e não-pagamento de dividendos.

A composição do endividamento também tem relevância na previsão de dificuldades financeiras pois demonstra a política de captação de recursos dentro das organizações, pois esse quociente revela em porcentagem que as empresas analisadas concentraram suas obrigações em curto prazo.

Projetos de investimentos e pagamento de dividendos concorrem pelas mesmas fontes de recursos, por isso a variável FATA - mostra a de rotatividade de ativos - reflete eficiência com que a empresa utiliza seus ativos, ou seja, mede especificamente a capacidade de uma empresa gerar vendas a partir de investimentos em ativos fixos. Em geral, uma taxa de rotatividade de ativos mais alta indica que uma empresa utilizou de forma mais eficiente o investimento em ativos fixos para gerar receita.

Figura 14 - Variáveis de maior peso identificadas pelo modelo Xgboost.



Fonte: Dados da pesquisa.

3.7 Considerações finais

O objetivo principal neste capítulo foi prever a probabilidade de identificar clientes adimplentes e inadimplentes usuários de cartão de crédito por meio de um modelo utilizado em aprendizagem de máquina: *Extreme Gradient Boosting* (Xgboost). Esse modelo é uma evolução de outros métodos - como o AdaBoost e Random Forest - e foi aplicado em estudos recentes com intuito de prever falências de bancos e escores de crédito. Entretanto, sua aplicação nesta base de dados é inédita. Como objetivo secundário, este estudo concentrou-se

no poder preditivo em vez de explorar a construção e desenvolvimento do modelo. Almejou, também, destacar os benefícios do uso dos algoritmos de aprendizagem de máquina aplicados às pesquisas de finanças.

O estudo mostrou que o Xgboost possui maior poder preditivo quando comparado a métodos mais populares, como: Regressão Logística e Random Forest, considerando os parâmetros utilizados. Ademais, o Xgboost possui um recurso importante, que corrobora com Chen e Li (2010), Wang e Ma (2011), Tsai, Hsu e Yen (2014); Zhao et al. (2015), onde sugerem que modelos e resultados deveriam se concentrar em dar explicações sobre as razões pelas quais empresas enfrentaram dificuldades financeiras; fatores que são importantes tanto para as corporações quanto para as instituições financeiras. Como não há uma resposta concreta sobre as características mais representativas (variáveis independentes), o método proposto demonstrou que a redução na distribuição dos dividendos classifica uma empresa com alta probabilidade de enfrentar um futuro aperto financeiro e seus investimentos também contribuem para isto.

A capacidade preditiva do modelo testado neste estudo deve encorajar novas pesquisas a unir forças com cientistas da computação e adicionar uma dinâmica aos modelos econométricos comumente adotados nos estudos em finanças. Além disso, em uma tentativa de estender os limites atuais de desempenho e interpretabilidade, o Xgboost rastreia variáveis que podem adicionar um peso preditivo extra ao modelo e trazer uma agilidade operacional a todo o processo.

Como limitação da pesquisa, destaca-se que parâmetros devem ser melhor explorados e outras técnicas, tal como Redes Neurais, poderiam contrapor os resultados encontrados. Porém, mesmo diante da limitação proveniente de possível endogeneidade do modelo, o estudo avança ao reforçar e estimular a automação dos processos e modelos utilizados. Por um lado, um amplo volume de diferentes dados pode ser extremamente benéfico para a aprendizagem das máquinas; por outro, precisa-se pensar em agilidade no processo de tomada de decisão e isso será alcançado com o maior número de informações relevantes sobre clientes, desenvolvimentos de algoritmos ágeis e que maximizem a capacidade operacional dos computadores disponíveis.

Em face dos resultados, acredita-se que as empresas possam desenvolver sistemas preventivos que alertariam seus gestores sobre indicadores que possam afetar sua saúde financeira. Além disso, as instituições poderiam evitar a inadimplência adotando medidas cautelares apropriadas, em vez de esperar até que a restrição financeira aconteça.

Em virtude dos fatos e resultados encontrados nesta pesquisa, este artigo tem interesse inconteste no aperfeiçoamento do modelo observando alguns aspectos, dentre eles estão:

- Fomentar a construção e expansão das bases de dados com diferentes variáveis, porém, relevantes à pesquisa de crédito que poderiam melhorar a robustez do modelo;
- Pesquisadores devem estruturar bases de dados que contenham variáveis diversas e relevantes à área – tanto informações financeiras, quanto aspectos qualitativos – e estimular sua adoção em instituições financeiras para aprimorar, assim, tanto o processo quanto as previsões de novos modelos;
- Avaliar a divisão ideal de uma base de dados para treinamento e teste do modelo;
- Experimentar diferentes parâmetros em diferentes dados, como variáveis corporativas ou informações de empresas de mercados emergentes;
- Incluir variáveis qualitativas (por exemplo, informações sociais, macroeconômicas, qualidade da gestão, etc.).

CAPITULO 4 – CONCLUSÃO

Corroborando com o diagnóstico divulgado por Mckinsey and Company (2015), os resultados encontrados no estudo mostram que há um cenário de viabilidade de estudos sobre o tema, o que demonstra que a discussão ainda pode ser considerada incipiente. O objetivo dessa pesquisa consistiu em contribuir com mais evidências empíricas ao demonstrar que o uso de sistemas inteligentes na gestão de risco de crédito tem ganhos incrementais, uma vez que a literatura tem apontado que a gestão de risco financeiro inteligente possui importantes implicações tanto para instituições financeiras, quanto para a economia de modo geral.

Para se atingir o objetivo do estudo, foi utilizado a metodologia bibliométrica Proknow-C para analisar a literatura disponível sobre o tema e chegar a um portfólio final com 168 estudos, considerados capazes de descrever o cenário da análise de crédito por meio do aprendizado de máquinas e apontar suas tendências no primeiro artigo e no segundo artigo foi utilizado o algoritmo Xgboost para verificar sua eficácia e capacidade de processamento de informações, comparado a modelos mais conhecidos, como: Regressão Logística e Árvore Aleatória. A escolha desse algoritmo deve-se às lacunas encontradas na revisão da literatura.

O modelo foi treinado e testado utilizando 1.055 empresas latino-americanas, cujas informações foram retiradas da base Economática. As variáveis independentes utilizadas no modelo são compostas de índices financeiros. A variável de interesse é definida por duas condições, quais são: 1) lucro antes de juros e impostos depreciação e amortização (EBITDA) são inferiores às suas despesas financeiras para dois anos consecutivos, (2) uma queda em seu valor de mercado ocorre entre dois períodos consecutivos (MANZANEQUE; MERINO; PRIEGO, 2016). Caso a empresa em questão satisfaça as duas condições, ela é classificada com dificuldades financeiras (valor 1), e sem dificuldades financeiras (valor 0).

Os resultados do primeiro estudo evidenciaram que o tema ainda carece de uma capacidade de mensurar variáveis qualitativas relativas ao comportamento da empresas e bases que possam conter características que representem o comportamento cíclico tanto das corporações quanto do mercado. Os resultados também sugerem que modelos de aprendizado de máquina devem apresentar uma acurácia confiável, considerando o volume de dados disponíveis e a velocidade em que esses são processados. De forma adicional, foi identificado que resultados apresentados pelos modelos dever-se-iam ser de fácil entendimento e a partir dessas respostas, criar sistemas preventivos tanto para as instituições quanto para os clientes, que possam indicar um limiar para o risco de credito antes que a inadimplência aconteça.

Com relação ao segundo artigo, dentre todas lacunas identificadas, o modelo Xgboost, conseguiu superar em capacidade de previsão, mesmo que de forma incremental o modelo Árvore Aleatória, um algoritmo utilizado em outras tarefas de aprendizado de máquina.

Mesmo diante das contribuições apontadas, destaca-se que o estudo possui algumas limitações. O período de análise da literatura poderia se estender com a intenção de verificar o *status quo* das pesquisas não só pós crise, mas anteriores a 2008.

Outra limitação refere-se a coleta de dados para estudo o qual ainda é um fator decisório que precisa ser de mais fácil acesso, posto que instituições financeiras oferecem dificuldades em dispor de dados de clientes. Sugere-se, então, o desenvolvimento e estruturação de base de dados específicas para o desenvolvimento de modelos de crédito.

Porém, mesmo diante das limitações existentes, o estudo avança ao demonstrar que a união entre o estudo de finanças e a ciência da computação - com sua automação de análises - pode ser considerada um elemento importante, quando se trata de estudos sobre precisão de análise e o volume de informações que são gerados diariamente sobre o comportamento do consumidor. Para futuras pesquisas, sugere-se a replicação da metodologia Proknow-C incorporando novas palavras-chave à fase inicial da pesquisa, bem como o período analisado. O uso de variáveis não financeiras para avaliação do risco de crédito pode constituir uma oportunidade de pesquisa futura.

Ainda como sugestão de exploração futura, sugere-se uma investigação mais minuciosa sobre os parâmetros utilizados para calibrar o algoritmo, bem como a melhor divisão dos dados- tanto para o treinamento quanto para o teste do modelo. Outra sugestão seria a realização de uma análise com dados qualitativos de empresas de mercados emergente com dados qualitativos, conforme estudos de Lin; Liang; Chen (2011) e Kim (2011), onde sugerem a construção de modelos usando fatores relacionados à governança corporativa, participação de mercado, estilo de gestão e perspectiva da indústria, etc.

REFERÊNCIAS

ABDOU, Hussein A. Genetic programming for credit scoring: The case of Egyptian public sector banks. Expert systems with applications, v. 36, n. 9, p. 11402-11417, 2009. https://doi.org/10.1016/j.eswa.2009.01.076

ADRIAANS, Pieter; ZANTINGE, Dolf. Data mining. Harlow: Addison-Wesley, 1996.

ALFARO, Esteban et al. Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks. Decision Support Systems, v. 45, n. 1, p. 110-122, 2008. https://doi.org/10.1016/j.dss.2007.12.002

ALP, Özge Sezgin et al. CMARS and GAM & CQP—modern optimization methods applied to international credit default prediction. Journal of computational and applied mathematics, v. 235, n. 16, p. 4639-4651, 2011. https://doi.org/10.1016/j.cam.2010.04.039

ALTMAN, Edward I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The journal of finance, v. 23, n. 4, p. 589-609, 1968. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

BAE, Jae Kwon. Predicting financial distress of the South Korean manufacturing industries. Expert Systems with Applications, v. 39, n. 10, p. 9159-9165, 2012. https://doi.org/10.1016/j.eswa.2012.02.058

BAESENS, B. M. M. Developing intelligent systems for credit scoring using machine learning techniques. 2004.

BANASIK, John; CROOK, Jonathan. Reject inference, augmentation, and sample selection. European Journal of Operational Research, v. 183, n. 3, p. 1582-1594, 2007. https://doi.org/10.1016/j.ejor.2006.06.072

BASEL COMMITTEE ON BANKING SUPERVISION (BIS). Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework - Comprehensive Version, Bank for International Settlements. 2006. Disponível em: https://www.bis.org/publ/bcbs118.pdf. Acesso em: 08/08/2017.

BATISTA, Gustavo EAPA; MONARD, Maria Carolina. An analysis of four missing data treatment methods for supervised learning. Applied artificial intelligence, v. 17, n. 5-6, p. 519-533, 2003. https://doi.org/10.1080/713827181

BENMELECH, Efraim; BERGMAN, Nittai K. Bankruptcy and the collateral channel. The Journal of Finance, v. 66, n. 2, p. 337-378, 2011. https://doi.org/10.1111/j.1540-6261.2010.01636.x

BIGUS, Joseph P. Data mining with neural networks: solving business problems from application development to decision support. 1996.

BLANCO, Antonio et al. Credit scoring models for the microfinance industry using neural networks: Evidence from Peru. Expert Systems with applications, v. 40, n. 1, p. 356-364, 2013. https://doi.org/10.1016/j.eswa.2012.07.051

BREIMAN, Leo. Random forests. Machine learning, v. 45, n. 1, p. 5-32, 2001. https://doi.org/10.1023/A:1010933404324

BROWN, Iain; MUES, Christophe. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, v. 39, n. 3, p. 3446-3453, 2012. https://doi.org/10.1016/j.eswa.2011.09.033

CAMPOS, A. L. S.; NAKAMURA, W. T. Rebalanceamento da Estrutura de Capital: endividamento setorial e folga financeira. Revista de Administração Contemporânea, v. 19,p. 20-37, 2015. https://doi.org/10.1590/1982-7849rac20151789

CARMONA, Pedro; CLIMENT, Francisco; MOMPARLER, Alexandre. Predicting failure in the US banking sector: An extreme gradient boosting approach. International Review of Economics & Finance, 2018. https://doi.org/10.1016/j.iref.2018.03.008

CHAWLA, Nitesh V. et al. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, v. 16, p. 321-357, 2002. https://doi.org/10.1613/jair.953

CHEN, Hui-Ling et al. A novel bankruptcy prediction model based on an adaptive fuzzy k-nearest neighbor method. Knowledge-Based Systems, v. 24, n. 8, p. 1348-1359, 2011. https://doi.org/10.1016/j.knosys.2011.06.008

CHEN, Mu-Chen; HUANG, Shih-Hsien. Credit scoring and rejected instances reassigning through evolutionary computation techniques. Expert Systems with Applications, v. 24, n. 4, p. 433-441, 2003. https://doi.org/10.1016/S0957-4174(02)00191-4

CHEN, Tianqi; GUESTRIN, Carlos. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM, p. 785-794, 2016. 10.1145/2939672.2939785

CHEN, Xiaofang et al. The Credit Scoring Model Based on Logistic-BP-AdaBoost Algorithm and its Application in P2P Credit Platform. In: Proceedings of the Fourth International Forum on Decision Sciences. Springer, Singapore, p. 119-130, 2017. https://doi.org/10.1007/978-981-10-2920-2_11

CHEN, Wei-Sen; DU, Yin-Kuan. Using neural networks and data mining techniques for the financial distress prediction model. Expert systems with applications, v. 36, n. 2, p. 4075-4086, 2009. https://doi.org/10.1016/j.eswa.2008.03.020

COUDERT, Virginie; GEX, Mathieu. Contagion inside the credit default swaps market: The case of the GM and Ford crisis in 2005. Journal of International Financial Markets, Institutions and Money, v. 20, n. 2, p. 109-134, 2010. https://doi.org/10.1016/j.intfin.2010.01.001

DESAI, Vijay S.; CROOK, Jonathan N.; OVERSTREET JR, George A. A comparison of neural networks and linear scoring models in the credit union environment. European Journal of Operational Research, v. 95, n. 1, p. 24-37, 1996. https://doi.org/10.1016/0377-2217(95)00246-4

DIRICK, Lore et al. Macro-economic factors in credit risk calculations: including time-varying covariates in mixture cure models. Journal of Business & Economic Statistics, p. 1-14, 2017.

DOMINGOS, Pedro. Metacost: A general method for making classifiers cost-sensitive. In: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, p. 155-164, 1999.

ENSSLIN, Leonardo et al. BPM governance: a literature analysis of performance evaluation. Business Process Management Journal, v. 23, n. 1, p. 71-86, 2017.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. AI magazine, v. 17, n. 3, p. 37, 1996. https://doi.org/10.1609/aimag.v17i3.1230

FINLAY, Steven. Multiple classifier architectures and their application to credit risk assessment. European Journal of Operational Research, v. 210, n. 2, p. 368-378, 2011. https://doi.org/10.1016/j.ejor.2010.09.029

FMI – FUNDO MONETÁRIO INTERNACIONAL. IMF Survey: Further Action Needed to Reinforce Signs of Market Recovery: IMF, 2009. Disponível em:<https://www.imf.org/en/News/Articles/2015/09/28/04/53/sores042109c>. Acesso em: 03/08/2017.

FRIEDMAN, Jerome H. Greedy function approximation: a gradient boosting machine. Annals of statistics, p. 1189-1232, 2001.

FRIEDMAN, Jerome H. Stochastic gradient boosting. Computational Statistics & Data Analysis, v. 38, n. 4, p. 367-378, 2002. https://doi.org/10.1016/S0167-9473(01)00065-2

FRIEDMAN, Jerome; HASTIE, Trevor; TIBSHIRANI, Robert. The elements of statistical learning. New York, NY, USA: Springer series in statistics, 2001.

FRYDMAN, Halina; ALTMAN, Edward I.; KAO, Duen-Li. Introducing recursive partitioning for financial classification: the case of financial distress. The Journal of Finance, v. 40, n. 1, p. 269-291, 1985. https://doi.org/10.1111/j.1540-6261.1985.tb04949.x

GALINDO, Jorge; TAMAYO, Pablo. Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications. Computational Economics, v. 15, n. 1-2, p. 107-143, 2000. https://doi.org/10.1023/A:1008699112516

GARFIELD, Eugene. The history and meaning of the journal impact factor. Jama, v. 295, n. 1, p. 90-93, 2006. https://doi.org/10.1001/jama.295.1.90

GUJARATI, Damodar N.; PORTER, Dawn C. Econometria básica. 5. ed. Porto Alegre: AMGH, 2011.

GUO, Yanhong et al. Instance-based credit risk assessment for investment decisions in P2P lending. European Journal of Operational Research, v. 249, n. 2, p. 417-426, 2016a. https://doi.org/10.1016/j.ejor.2015.05.050

GUO, Guangming et al. From footprint to evidence: An exploratory study of mining social data for credit scoring. ACM Transactions on the Web (TWEB), v. 10, n. 4, p. 22, 2016b.

HE, Hongliang; ZHANG, Wenyu; ZHANG, Shuai. A novel ensemble method for credit scoring: Adaption of different imbalance ratios. Expert Systems with Applications, v. 98, p. 105-117, 2018. https://doi.org/10.1016/j.eswa.2018.01.012

HE, T. An Introduction to XGBoost R package. Distributed Machine Learning Community. 2016. Disponível em: http://dmlc.ml/rstats/2016/03/10/XGBoost.html. Acesso em: 09/07/2018.

HEO, Junyoung; YANG, Jin Yong. AdaBoost based bankruptcy forecasting of Korean construction companies. Applied soft computing, v. 24, p. 494-499, 2014. https://doi.org/10.1016/j.asoc.2014.08.009

HERTZEL, Michael G.; OFFICER, Micah S. Industry contagion in loan spreads. Journal of Financial Economics, v. 103, n. 3, p. 493-506, 2012. https://doi.org/10.1016/j.jfineco.2011.10.012

HUANG, Jih-Jeng; TZENG, Gwo-Hshiung; ONG, Chorng-Shyong. Two-stage genetic programming (2SGP) for the credit scoring model. Applied Mathematics and Computation, v. 174, n. 2, p. 1039-1053, 2006. https://doi.org/10.1016/j.amc.2005.05.027

HUANG, Zan et al. Credit rating analysis with support vector machines and neural networks: a market comparative study. Decision support systems, v. 37, n. 4, p. 543-558, 2004. https://doi.org/10.1016/S0167-9236(03)00086-1

IQUIAPAZA, Robert Aldo; LAMOUNIER, Wagner Moura; AMARAL, Hudson Fernandes. Assimetric information and dividends payout at the Sao Paulo stock exchange (Bovespa). Ad. Sci. appl. Account, 2008.

JAPKOWICZ, Nathalie. The class imbalance problem: Significance and strategies. In: Proc. of the Int'l Conf. on Artificial Intelligence. 2000.

JONES, Stewart; JOHNSTONE, David; WILSON, Roy. An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes. Journal of Banking & Finance, v. 56, p. 72-85, 2015. https://doi.org/10.1016/j.jbankfin.2015.02.006

KHASHMAN, Adnan. Credit risk evaluation using neural networks: Emotional versus conventional models. Applied Soft Computing, v. 11, n. 8, p. 5477-5484, 2011. https://doi.org/10.1016/j.asoc.2011.05.011

KHASHMAN, Adnan. Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. Expert Systems with Applications, v. 37, n. 9, p. 6233-6239, 2010. https://doi.org/10.1016/j.eswa.2010.02.101

KIM, Myoung-Jong; KANG, Dae-Ki. Ensemble with neural networks for bankruptcy prediction. Expert systems with applications, v. 37, n. 4, p. 3373-3379, 2010. https://doi.org/10.1016/j.eswa.2009.10.012

KIM, Myoung-Jong; KANG, Dae-Ki; KIM, Hong Bae. Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction. Expert Systems with Applications, v. 42, n. 3, p. 1074-1082, 2015. https://doi.org/10.1016/j.eswa.2014.08.025

KIM, Soo Y. Prediction of hotel bankruptcy using support vector machine, artificial neural network, logistic regression, and multivariate discriminant analysis. The Service Industries Journal, v. 31, n. 3, p. 441-468, 2011. https://doi.org/10.1080/02642060802712848

KRUGMAN, Paul. The Return of Depression Economics and the Crisis of 2008. Nova Iorque: W. W. Norton & Company, 2008.

LAI, Kin Keung et al. Credit risk analysis using a reliability-based neural network ensemble model. In: International Conference on Artificial Neural Networks. Springer, Berlin, Heidelberg, p. 682-690, 2006. https://doi.org/10.1007/11840930_71

LEE, Tian-Shyug et al. Credit scoring using the hybrid neural discriminant technique. Expert Systems with applications, v. 23, n. 3, p. 245-254, 2002. https://doi.org/10.1016/S0957-4174(02)00044-1

LEE, Tian-Shyug; CHEN, I.-Fei. A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. Expert Systems with Applications, v. 28, n. 4, p. 743-752, 2005. https://doi.org/10.1016/j.eswa.2004.12.031

LENSBERG, Terje; EILIFSEN, Aasmund; MCKEE, Thomas E. Bankruptcy theory development and classification via genetic programming. European Journal of Operational Research, v. 169, n. 2, p. 677-697, 2006. https://doi.org/10.1016/j.ejor.2004.06.013

LEONARD, Kevin J. Information systems and benchmarking in the credit scoring industry. Benchmarking for Quality Management & Technology, v. 3, n. 1, p. 38-44, 1996. https://doi.org/10.1108/14635779610112449

LI, Hui et al. The random subspace binary logit (RSBL) model for bankruptcy prediction. Knowledge-Based Systems, v. 24, n. 8, p. 1380-1388, 2011a. https://doi.org/10.1016/j.knosys.2011.06.015

LI, Hui; SUN, Jie. Empirical research of hybridizing principal component analysis with multivariate discriminant analysis and logistic regression for business failure prediction. Expert Systems with Applications, v. 38, n. 5, p. 6244-6253, 2011b. https://doi.org/10.1016/j.eswa.2010.11.043

LI, Hui; SUN, Jie. Majority voting combination of multiple case-based reasoning for financial distress prediction. Expert Systems with Applications, v. 36, n. 3, p. 4363-4373, 2009. https://doi.org/10.1016/j.eswa.2008.05.019

LI, Jianping et al. Evolution strategy based adaptive Lq penalty support vector machines with Gauss kernel for credit risk analysis. Applied Soft Computing, v. 12, n. 8, p. 2675-2682, 2012. https://doi.org/10.1016/j.asoc.2012.04.011

LI, Zhiyong et al. Reject inference in credit scoring using Semi-supervised Support Vector Machines. Expert Systems with Applications, v. 74, p. 105-114, 2017. https://doi.org/10.1016/j.eswa.2017.01.011

LIANG, Deron et al. Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. European Journal of Operational Research, v. 252, n. 2, p. 561-572, 2016. https://doi.org/10.1016/j.ejor.2016.01.012

LIN, Fengyi; LIANG, Deron; CHEN, Enchia. Financial ratio selection for business crisis prediction. Expert Systems with Applications, v. 38, n. 12, p. 15094-15102, 2011. https://doi.org/10.1016/j.eswa.2011.05.035

LIU, Ling; ÖZSU, M. Tamer. Encyclopedia of database systems. New York, NY, USA: Springer, 2009.

LOTERMAN, Gert et al. Benchmarking regression algorithms for loss given default modeling. International Journal of Forecasting, v. 28, n. 1, p. 161-170, 2012. https://doi.org/10.1016/j.ijforecast.2011.01.006

MALHOTRA, Rashmi; MALHOTRA, Davinder K. Differentiating between good credits and bad credits using neuro-fuzzy systems. European journal of operational research, v. 136, n. 1, p. 190-211, 2002. https://doi.org/10.1016/S0377-2217(01)00052-2

MALHOTRA, Rashmi; MALHOTRA, Davinder K. Evaluating consumer loans using neural networks. Omega, v. 31, n. 2, p. 83-96, 2003. https://doi.org/10.1016/S0305-0483(03)00016-1

MANZANEQUE, Montserrat; MERINO, Elena; PRIEGO, Alba María. The role of institutional shareholders as owners and directors and the financial distress likelihood. Evidence from a concentrated ownership context. European Management Journal, v. 34, n. 4, p. 439-451, 2016. https://doi.org/10.1016/j.emj.2016.01.007

MARKOVITCH, Shaul; ROSENSTEIN, Dan. Feature generation using general constructor functions. Machine Learning, v. 49, n. 1, p. 59-98, 2002. https://doi.org/10.1023/A:1014046307775

McKINSEY and COMPANY. The future of bank risk management, 2015. Disponível em: https://www.mckinsey.com/business-functions/risk/our-insights/the-future-of-bank-risk-management. Acesso em: 03/08/2017.

Microsoft Azure Machine Learning Studio. Disponível em: < https://studio.azureml.net/> Acesso em: 10/01/2019.

OHLSON, James A. Financial ratios and the probabilistic prediction of bankruptcy. Journal of accounting research, p. 109-131, 1980.

ONG, Chorng-Shyong; HUANG, Jih-Jeng; TZENG, Gwo-Hshiung. Building credit scoring models using genetic programming. Expert Systems with Applications, v. 29, n. 1, p. 41-47, 2005. https://doi.org/10.1016/j.eswa.2005.01.003

OPLER, Tim C.; TITMAN, Sheridan. Financial distress and corporate performance. The Journal of Finance, v. 49, n. 3, p. 1015-1040, 1994. https://doi.org/10.1111/j.1540-6261.1994.tb00086.x

PALIWAL, Mukta; KUMAR, Usha A. Neural networks and statistical techniques: A review of applications. Expert systems with applications, v. 36, n. 1, p. 2-17, 2009. https://doi.org/10.1016/j.eswa.2007.10.005

PAZZANI, Michael et al. Reducing misclassification costs. In: Machine Learning Proceedings 1994. 1994. p. 217-225. https://doi.org/10.1016/B978-1-55860-335-6.50034-9

PINDADO, Julio; RODRIGUES, Luis; DE LA TORRE, Chabela. Estimating financial distress likelihood. Journal of Business Research, v. 61, n. 9, p. 995-1003, 2008. https://doi.org/10.1016/j.jbusres.2007.10.006

PIRAMUTHU, Selwyn. Financial credit-risk evaluation with neural and neurofuzzy systems. European Journal of Operational Research, v. 112, n. 2, p. 310-321, 1999. https://doi.org/10.1016/S0377-2217(97)00398-6

SAMUEL, Arthur L. Some studies in machine learning using the game of checkers. IBM Journal of research and development, v. 3, n. 3, p. 210-229, 1959. https://doi.org/10.1147/rd.33.0210

SANZ, Luis J.; AYCA, Julio. Financial distress costs in Latin America: A case study. Journal of Business Research, v. 59, n. 3, p. 394-395, 2006. https://doi.org/10.1016/j.jbusres.2005.09.014

SCHRICKEL, Wolfgang Kurt. Análise de crédito: concessão e gerência de empréstimos. 5. ed. São Paulo: Atlas, 2000.

SCIMAGO JOURNAL RANK (SJR), Journal Rankings Disponível em: < https://www.scimagojr.com/journalrank.php> Acesso em: 28/08/2017.

SILVA DA ROSA, Fabricia et al. Environmental disclosure management: a constructivist case. Management Decision, v. 50, n. 6, p. 1117-1136, 2012. https://doi.org/10.1108/00251741211238364

SMALZ, Robert; CONRAD, Michael. Combining evolution with credit apportionment: A new learning algorithm for neural nets. Neural Networks, v. 7, n. 2, p. 341-351, 1994. https://doi.org/10.1016/0893-6080(94)90028-0

SUN, Jie et al. Dynamic financial distress prediction with concept drift based on time weighting combined with Adaboost support vector machine ensemble. Knowledge-Based Systems, v. 120, p. 4-14, 2017. https://doi.org/10.1016/j.knosys.2016.12.019

SUN, Jie; JIA, Ming-Yue; LI, Hui. AdaBoost ensemble for financial distress prediction: An empirical comparison with data from Chinese listed companies. Expert Systems with Applications, v. 38, n. 8, p. 9305-9312, 2011. https://doi.org/10.1016/j.eswa.2011.01.042

SUN, Jie; LI, Hui. Financial distress prediction based on serial combination of multiple classifiers. Expert Systems with Applications, v. 36, n. 4, p. 8659-8666, 2009. https://doi.org/10.1016/j.eswa.2008.10.002

THOMAS, Lyn C. Methodologies for classifying applicants for credit. Statistics in finance, p. 83-103, 1998.

THOMSON REUTERS, Journal citation reports (JCR), Disponível em: < https://www.thomsonreuters.com.br/pt.html > Acesso em: 27/08/2017.

TRANFIELD, David; DENYER, David; SMART, Palminder. Towards a methodology for developing evidence-informed management knowledge by means of systematic review. British journal of management, v. 14, n. 3, p. 207-222, 2003. https://doi.org/10.1111/1467-8551.00375

TSAI, Chih-Fong; HSU, Yu-Feng; YEN, David C. A comparative study of classifier ensembles for bankruptcy prediction. Applied Soft Computing, v. 24, p. 977-984, 2014. https://doi.org/10.1016/j.asoc.2014.08.047

TSENG, Fang-Mei; HU, Yi-Chung. Comparing four bankruptcy prediction models: Logit, quadratic interval logit, neural and fuzzy neural networks. Expert Systems with Applications, v. 37, n. 3, p. 1846-1853, 2010. https://doi.org/10.1016/j.eswa.2009.07.081

VAN GESTEL, Ir Tony et al. A support vector machine approach to credit scoring. In: Forum Financier-Revue Bancaire Et Financiaire Bank En Financiewezen-. Unknown, p. 73-82, 2003.

VAN GESTEL, Tony et al. Bayesian kernel based classification for financial distress detection. European journal of operational research, v. 172, n. 3, p. 979-1003, 2006. https://doi.org/10.1016/j.ejor.2004.11.009

VARETTO, Franco. Genetic algorithms applications in the analysis of insolvency risk. Journal of Banking & Finance, v. 22, n. 10-11, p. 1421-1439, 1998. https://doi.org/10.1016/S0378-4266(98)00059-4

XIA, Yufei; YANG, Xiaoli; ZHANG, Yeying. A Rejection Inference Technique Based on Contrastive Pessimistic Likelihood Estimation for P2P Lending. Electronic Commerce Research and Applications, 2018. https://doi.org/10.1016/j.elerap.2018.05.011

WANG, Gang et al. A comparative assessment of ensemble learning for credit scoring. Expert systems with applications, v. 38, n. 1, p. 223-230, 2011. https://doi.org/10.1016/j.eswa.2010.06.048

WANG, Gang; MA, Jian. Study of corporate credit risk prediction based on integrating boosting and random subspace. Expert Systems with Applications, v. 38, n. 11, p. 13871-13878, 2011. https://doi.org/10.1016/j.eswa.2011.04.191

WANG, Gang; MA, Jian; YANG, Shanlin. An improved boosting based on feature selection for corporate bankruptcy prediction. Expert Systems with Applications, v. 41, n. 5, p. 2353-2361, 2014. https://doi.org/10.1016/j.eswa.2013.09.033

WANG, Jian; VEUGELERS, Reinhilde; STEPHAN, Paula. Bias against novelty in science: A cautionary tale for users of bibliometric indicators. Research Policy, v. 46, n. 8, p. 1416-1436, 2017. https://doi.org/10.1016/j.respol.2017.06.006

WANG, Yongqiao; WANG, Shouyang; LAI, Kin Keung. A new fuzzy support vector machine to evaluate credit risk. IEEE Transactions on Fuzzy Systems, v. 13, n. 6, p. 820-831, 2005. https://doi.org/10.1109/TFUZZ.2005.859320

WAQAS, Hamid; MD-RUS, Rohani. Predicting financial distress: Importance of accounting and firm-specific market variables for Pakistan's listed firms. Cogent Economics & Finance, v. 6, n. 1, p. 1-16, 2018. https://doi.org/10.1080/23322039.2018.1545739

WEST, David. Neural network credit scoring models. Computers & Operations Research, v. 27, n. 11-12, p. 1131-1152, 2000. https://doi.org/10.1016/S0305-0548(99)00149-5

XIA, Yufei 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. https://doi.org/10.1016/j.eswa.2017.02.017

XIA, Yufei et al. A novel heterogeneous ensemble credit-scoring model based on bstacking approach. Expert Systems with Applications, v. 93, p. 182-199, 2018. https://doi.org/10.1016/j.eswa.2017.10.022

YEH, Ching-Chiang; CHI, Der-Jang; HSU, Ming-Fu. A hybrid approach of DEA, rough set and support vector machines for business failure prediction. Expert Systems with Applications, v. 37, n. 2, p. 1535-1541, 2010. https://doi.org/10.1016/j.eswa.2009.06.088

YU, Lean et al. An ELM-based Classification Algorithm with Optimal Cutoff Selection for Credit Risk Assessment. Filomat, v. 30, n. 15, p. 4027-4036, 2016. https://doi.org/10.2298/FIL1615027Y

YU, Lei; LIU, Huan. Efficient feature selection via analysis of relevance and redundancy. Journal of machine learning research, v. 5, n. Oct, p. 1205-1224, 2004.

ZHANG, Defu et al. Vertical bagging decision trees model for credit scoring. Expert Systems with Applications, v. 37, n. 12, p. 7838-7843, 2010. https://doi.org/10.1016/j.eswa.2010.04.054

ZHANG, Lili; PRIESTLEY, Jennifer; NI, Xuelei. Influence of the Event Rate on Discrimination Abilities of Bankruptcy Prediction Models. arXiv preprint arXiv:1803.03756, 2018.

ZHANG, Shichao; ZHANG, Chengqi; YANG, Qiang. Data preparation for data mining. Applied artificial intelligence, v. 17, n. 5-6, p. 375-381, 2003. https://doi.org/10.1080/713827180

ZHAO, Zongyuan et al. Investigation and improvement of multi-layer perceptron neural networks for credit scoring. Expert Systems with Applications, v. 42, n. 7, p. 3508-3516, 2015. https://doi.org/10.1016/j.eswa.2014.12.006

APÊNDICE I – Portfólio Final contendo 168 artigos selecionados via Proknow-C.

Titulo	Autores	Ano
Using neural network ensembles for bankruptcy prediction and credit scoring	Tsai, C. F.; Wu, J. W.	2008
Financial distress and corporate risk management: Theory and evidence	Purnanandam, Amiyatosh	2008
Credit risk assessment with a multistage neural network ensemble learning approach	Yu, L.; Wang, S. Y.; Lai, K. K.	2008
A neural network approach for credit risk evaluation	Angelini, Eliana; di Tollo, Giacomo; Roli, Andrea	2008
Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks	Alfaro, E.; Garcia, N.; Gamez, M.; Elizondo, D.	2008
Data mining method for listed companies' financial distress prediction	Sun, J.; Li, H.	2008
Threshold accepting trained principal component neural network and feature subset selection: Application to bankruptcy prediction in banks	Ravi, V.; Pramodh, C.	2008
Feature selection to diagnose a business crisis by using a real GA-based support vector machine: An empirical study	Chen, L. H.; Hsiao, H. D.	2008
A hybrid financial analysis model for business failure prediction	Huang, S. M.; Tsai, C. F.; Yen, D. C.; Cheng, Y. L.	2008
Optimization of nearest neighbor classifiers via metaheuristic algorithms for credit risk assessment	Marinakis, Y.; Marinaki, M.; Doumpos, M.; Matsatsinis, N.; Zopounidis, C.	2008
Using neural networks and data mining techniques for the financial distress prediction model	Chen, W. S.; Du, Y. K.	2009
Feature selection in bankruptcy prediction	Tsai, C. F.	2009
Support vector machines for credit scoring and discovery of significant features	Bellotti, T.; Crook, J.	2009
An intelligent-agent-based fuzzy group decision making model for financial multicriteria decision support: The case of credit scoring	Yu, L.; Wang, S. Y.; Lai, K. K.	2009
Differential evolution trained wavelet neural networks: Application to bankruptcy prediction in banks	Chauhan, N.; Ravi, V.; Chandra, D. K.	2009
Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach	Ahn, H.; Kim, K. J.	2009
An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring	Nanni, L.; Lumini, A.	2009
Consumer credit scoring models with limited data	Sustersic, M.; Mramor, D.; Zupan, J.	2009
A selective ensemble based on expected probabilities for bankruptcy prediction	Hung, C.; Chen, J. H.	2009
Genetic programming for credit scoring: The case of Egyptian public sector banks	Abdou, H. A.	2009
Random survival forests models for SME credit risk measurement	Fantazzini, D.; Figini, S.	2009
Financial distress prediction based on OR-CBR in the principle of k-nearest neighbors	Li, H.; Sun, J.; Sun, B. L.	2009
Gaussian case-based reasoning for business failure prediction with empirical data in China	Li, H.; Sun, J.	2009
A binary classification method for bankruptcy prediction	Min, Jae H.; Jeong, Chulwoo	2009
Developing a business failure prediction model via RST, GRA and CBR	Lin, R. H.; Wang, Y. T.; Wu, C. H.; Chuang, C. L.	2009
An integrative model with subject weight based on neural network learning for bankruptcy prediction	Cho, Sungbin; Kim, Jinhwa; Bae, Jae Kwon	2009
Credit scoring algorithm based on link analysis ranking with support vector machine	Xu, X.; Zhou, C.; Wang, Z.	2009

Majority voting combination of multiple case-based reasoning for financial distress prediction	Li, H.; Sun, J.	2009
Variable selection and oversampling in the use of smooth support vector machines for predicting the default risk of companies	Härdle, W.; Lee, Y. J.; Schäfer, D.; Yeh, Y. R.	2009
Prediction model building with clustering-launched classification and support vector machines in credit scoring	Härdle, W.; Lee, Y. J.; Schäfer, D.; Yeh, Y. R.Härdle, W.; Lee, Y. J.; Schäfer, D.; Yeh, Y. R.	2009
Alternative diagnosis of corporate bankruptcy: A neuro fuzzy approach	Chen, Hsueh-Ju; Huang, Shaio Yan; Lin, Chin-Shien	2009
A Comparison Of Data Mining Techniques For Credit Scoring In Banking: A Managerial Perspective	Ince, H.; Aktan, B.	2009
Financial distress prediction based on serial combination of multiple classifiers	Sun, J.; Li, H.	2009
Credit Scoring Models With AUC Maximization Based On Weighted SVM	Zhou, L. G.; Lai, K. K.; Yen, J.	2009
Accuracy of machine learning models versus "hand crafted" expert systems - A credit scoring case study	Ben-David, A.; Frank, E.	2009
Bankruptcy prediction using ELECTRE-based single-layer perceptron	Hu, Y. C.	2009
Are we modelling the right thing? The impact of incorrect problem specification in credit scoring	Finlay, S.	2009
Ensemble with neural networks for bankruptcy prediction	Kim, M. J.; Kang, D. K.	2010
A hybrid approach of DEA, rough set and support vector machines for business failure prediction	Yeh, C. C.; Chi, D. J.; Hsu, M. F.	2010
A data driven ensemble classifier for credit scoring analysis	Hsieh, N. C.; Hung, L. P.	2010
Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes	Khashman, A.	2010
Combination of feature selection approaches with SVM in credit scoring	Chen, F. L.; Li, F. C.	2010
Least squares support vector machines ensemble models for credit scoring	Zhou, Ligang; Lai, Kin Keung; Yu, Lean	2010
Support vector machine based multiagent ensemble learning for credit risk evaluation	Yu, L. A.; Yue, W. Y.; Wang, S. Y.; Lai, K. K.	2010
Subagging for credit scoring models	Paleologo, Giuseppe; Elisseeff, André; Antonini, Gianluca	2010
Multiple classifier application to credit risk assessment	Twala, B.	2010
Consumer credit-risk models via machine-learning algorithms	Khandani, A. E.; Kim, A. J.; Lo, A. W.	2010
Beyond business failure prediction	Wu, W. W.	2010
Predicting bankruptcy using neural networks and other classification methods: The influence of variable selection techniques on model accuracy	du Jardin, P.	2010
Predicting business failure using classification and regression tree: An empirical comparison with popular classical statistical methods and top classification mining methods	Li, H.; Sun, J.; Wu, J.	2010
A hybrid approach based on the combination of variable selection using decision trees and case-based reasoning using the Mahalanobis distance: For bankruptcy prediction	Cho, S.; Hong, H.; Ha, B. C.	2010
Vertical bagging decision trees model for credit scoring	Zhang, D.; Zhou, X.; Leung, S. C. H.; Zheng, J.	2010
Business failure prediction using decision trees	Gepp, A.; Kumar, K.; Bhattacharya, S.	2010
Financial distress prediction in banks using Group Method of Data Handling neural network, counter propagation neural network and fuzzy ARTMAP	Ravisankar, P.; Ravi, V.	2010
Contagion inside the credit default swaps market: The case of the GM and Ford crisis in 2005	Coudert, Virginie; Gex, Mathieu	2010

Business failure prediction using hybrid2 case-based reasoning (H2CBR)	Li, H.; Sun, J.	2010
Domain-driven classification based on multiple criteria and multiple constraint-level programming for intelligent credit scoring	He, J.; Zhang, Y.; Shi, Y.; Huang, G.	2010
Forecasting Business Failure in China Using Case-Based Reasoning with Hybrid Case Respresentation	Li, H.; Sun, J.	2010
A practical approach to bankruptcy prediction for small businesses: Substituting the unavailable financial data for credit card sales information	Yoon, J. S.; Kwon, Y. S.	2010
On sensitivity of case-based reasoning to optimal feature subsets in business failure prediction	Li, H.; Huang, H. B.; Sun, J.; Lin, C.	2010
From linear to non-linear kernel based classifiers for bankruptcy prediction	Van Gestel, T.; Baesens, B.; Martens, D.	2010
Support vector machine and wavelet neural network hybrid: Application to bankruptcy prediction in banks	Chandra, D. K.; Ravi, V.; Ravisankar, P.	2010
A comparative assessment of ensemble learning for credit scoring	Wang, Gang; Hao, Jinxing; Ma, Jian; Jiang, Hongbing	2011
Fuzzy Support Vector Machine for bankruptcy prediction	Chaudhuri, A.; De, K.	2011
A novel bankruptcy prediction model based on an adaptive fuzzy k-nearest neighbor method	Chen, H. L.; Yang, B.; Wang, G.; Liu, J.; Xu, X.; Wang, S. J.; Liu, D. Y.	2011
Using data mining to improve assessment of credit worthiness via credit scoring models	Yap, B. W.; Ong, S. H.; Husain, N. H. M.	2011
Bankruptcy forecasting: A hybrid approach using Fuzzy c-means clustering and Multivariate Adaptive Regression Splines (MARS)	De Andres, J.; Lorca, P.; Juez, F. J. D.; Sanchez-Lasheras, F.	2011
An empirical study of classification algorithm evaluation for financial risk prediction	Peng, Y.; Wang, G. X.; Kou, G.; Shi, Y.	2011
Multiple classifier architectures and their application to credit risk assessment	Finlay, S.	2011
Hybridizing principles of TOPSIS with case-based reasoning for business failure prediction	Li, H.; Adeli, H.; Sun, J.; Han, J. G.	2011
Hybrid neural intelligent system to predict business failure in small-to-medium-size enterprises	Borrajo, M. L.; Baruque, B.; Corchado, E.; Bajo, J.; Corchado, J. M.	2011
Credit risk evaluation using a weighted least squares SVM classifier with design of experiment for parameter selection	Yu, L.; Yao, X.; Wang, S.; Lai, K. K.	2011
The use of hybrid manifold learning and support vector machines in the prediction of business failure	Lin, F.; Yeh, C. C.; Lee, M. Y.	2011
Using partial least squares and support vector machines for bankruptcy prediction	Yang, Z.; You, W.; Ji, G.	2011
Financial ratio selection for business crisis prediction	Lin, F. Y.; Liang, D. R.; Chen, E. C.	2011
AdaBoost ensemble for financial distress prediction: An empirical comparison with data from Chinese listed companies	Sun, J.; Jia, M. Y.; Li, H.	2011
Predicting corporate bankruptcy using a self-organizing map: An empirical study to improve the forecasting horizon of a financial failure model	Du Jardin, P.; Séverin, E.	2011
Credit risk evaluation using neural networks: Emotional versus conventional models	Khashman, A.	2011
Prediction of hotel bankruptcy using support vector machine, artificial neural network, logistic regression, and multivariate discriminant analysis	Kim, S. Y.	2011
Dynamic financial distress prediction using instance selection for the disposal of concept drift	Sun, J.; Li, H.	2011
The random subspace binary logit (RSBL) model for bankruptcy prediction	Li, H.; Lee, Y. C.; Zhou, Y. C.; Sun, J.	2011
Principal component case-based reasoning ensemble for business failure prediction	Li, H.; Sun, J.	2011
A genetic algorithm-based approach to cost-sensitive bankruptcy prediction	Chen, N.; Ribeiro, B.; Vieira, A. S.; Duarte, J.; Neves, J. C.	2011

Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches	Chen, M. Y.	2011
A hybrid neural network approach for credit scoring	Chuang, C. L.; Huang, S. T.	2011
An enforced support vector machine model for construction contractor default prediction	Tserng, H. Ping; Lin, Gwo-Fong; Tsai, L. Ken; Chen, Po-Cheng	2011
Credit Risk Evaluation Model Development Using Support Vector Based Classifiers	Danenas, Paulius; Garsva, Gintautas; Gudas, Saulius	2011
Empirical research of hybridizing principal component analysis with multivariate discriminant analysis and logistic regression for business failure prediction	Li, H.; Sun, J.	2011
Artificial metaplasticity neural network applied to credit scoring	Marcano-Cedeño, A.; Marin-De-La-Barcena, A.; Jimenez-Trillo, J.; Piñuela, J. A.; Andina, D.	2011
Modeling default risk with support vector machines	Chen, S.; Härdle, W. K.; Moro, R. A.	2011
Study of corporate credit risk prediction based on integrating boosting and random subspace	Wang, G.; Ma, J.	2011
Credit risk estimation model development process: Main steps and model improvement	Mileris, R.; Boguslauskas, V.	2011
On performance of case-based reasoning in Chinese business failure prediction from sensitivity, specificity, positive and negative values	Li, H.; Sun, J.	2011
A new Fruit Fly Optimization Algorithm: Taking the financial distress model as an example	Pan, W. T.	2012
An experimental comparison of classification algorithms for imbalanced credit scoring data sets	Brown, Iain; Mues, Christophe	2012
Comparative analysis of data mining methods for bankruptcy prediction	Olson, D. L.; Delen, D.; Meng, Y. Y.	2012
An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data	Akkoç, S.	2012
Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment	Oreski, S.; Oreski, D.; Oreski, G.	2012
Two credit scoring models based on dual strategy ensemble trees	Wang, G.; Ma, J.; Huang, L. H.; Xu, K. Q.	2012
Benchmarking regression algorithms for loss given default modeling	Loterman, G.; Brown, I.; Martens, D.; Mues, C.; Baesens, B.	2012
Financial early warning system model and data mining application for risk detection	Koyuncugil, A. S.; Ozgulbas, N.	2012
Instance sampling in credit scoring: An empirical study of sample size and balancing	Crone, S. F.; Finlay, S.	2012
Financial distress prediction using support vector machines: Ensemble vs. individual	Sun, J.; Li, H.	2012
The prediction for listed companies' financial distress by using multiple prediction methods with rough set and Dempster-Shafer evidence theory	Xiao, Z.; Yang, X.; Pang, Y.; Dang, X.	2012
A hybrid ensemble approach for enterprise credit risk assessment based on Support Vector Machine	Wang, G.; Ma, J.	2012
Two-level classifier ensembles for credit risk assessment	Marqués, A. I.; García, V.; Sánchez, J. S.	2012
A hybrid approach to integrate genetic algorithm into dual scoring model in enhancing the performance of credit scoring model	Chi, B. W.; Hsu, C. C.	2012
An artificial immune classifier for credit scoring analysis	Chang, S. Y.; Yeh, T. Y.	2012
Computational time reduction for credit scoring: An integrated approach based on support vector machine and stratified sampling method	Hens, A. B.; Tiwari, M. K.	2012
Forecasting business failure: The use of nearest-neighbour support vectors and correcting imbalanced samples - Evidence from the Chinese hotel industry	Li, H.; Sun, J.	2012
A case-based reasoning model that uses preference theory functions for credit scoring	Vukovic, S.; Delibasic, B.; Uzelac, A.; Suknovic, M.	2012

A hybrid device for the solution of sampling bias problems in the forecasting of firms' bankruptcy	Sánchez-Lasheras, F.; De Andrés, J.; Lorca, P.; De Cos Juez, F. J.	2012
Classifiers selection in ensembles using genetic algorithms for bankruptcy prediction	Kim, M. J.; Kang, D. K.	2012
Exploring the behaviour of base classifiers in credit scoring ensembles	Marques, A. I.; Garcia, V.; Sanchez, J. S.	2012
Rough set and scatter search metaheuristic based feature selection for credit scoring	Wang, J.; Hedar, A. R.; Wang, S. Y.; Ma, J.	2012
Mining financial distress trend data using penalty guided support vector machines based on hybrid of particle swarm optimization and artificial bee colony algorithm	Hsieh, T. J.; Hsiao, H. F.; Yeh, W. C.	2012
Credit risk assessment and decision making by a fusion approach	Wu, T. C.; Hsu, M. F.	2012
Predicting financial distress of the South Korean manufacturing industries	Bae, J. K.	2012
Simple instance selection for bankruptcy prediction	Tsai, C. F.; Cheng, K. C.	2012
Credit risk Evaluation by hybrid data mining technique	Chen, Weimin; Xiang, Guocheng; Liu, Youjin; Wang, Kexi	2012
Does segmentation always improve model performance in credit scoring?	Bijak, Katarzyna; Thomas, Lyn C.	2012
A tuning method for the architecture of neural network models incorporating GAM and GA as applied to bankruptcy prediction	Jeong, C.; Min, J. H.; Kim, M. S.	2012
Bankruptcy prediction models based on multinorm analysis: An alternative to accounting ratios	De Andrés, J.; Landajo, M.; Lorca, P.	2012
Credit scoring models for the microfinance industry using neural networks: Evidence from Peru	Blanco, Antonio; Pino-Mejías, Rafael; Lara, Juan; Rayo, Salvador	2013
A multi-industry bankruptcy prediction model using back-propagation neural network and multivariate discriminant analysis	Lee, Sangjae; Choi, Wu Sung	2013
Clustering and visualization of bankruptcy trajectory using self-organizing map	Chen, N.; Ribeiro, B.; Vieira, A.; Chen, A.	2013
Consumer credit risk: Individual probability estimates using machine learning	Kruppa, J.; Schwarz, A.; Arminger, G.; Ziegler, A.	2013
On the suitability of resampling techniques for the class imbalance problem in credit scoring	Marqués, A. I.; García, V.; Sánchez, J. S.	2013
Partial Least Square Discriminant Analysis for bankruptcy prediction	Serrano-Cinca, C.; Gutierrez-Nieto, B.	2013
Bankruptcy prediction for Russian companies: Application of combined classifiers	Fedorova, Elena; Gilenko, Evgenii; Dovzhenko, Sergey	2013
A Rule-Based Model for Bankruptcy Prediction Based on an Improved Genetic Ant Colony Algorithm	Zhang, Y. D.; Wang, S. H.; Ji, G. L.	2013
A granular computing-based approach to credit scoring modeling	Saberi, M.; Mirtalaie, M. S.; Hussain, F. K.; Azadeh, A.; Hussain, O. K.; Ashjari, B.	2013
Improving the management of microfinance institutions by using credit scoring models based on Statistical Learning techniques	Cubiles-De-La-Vega, María-Dolores; Blanco-Oliver, Antonio; Pino-Mejías, Rafael; Lara-Rubio, Juan	2013
Evaluation of clustering algorithms for financial risk analysis using MCDM methods	Kou, G.; Peng, Y.; Wang, G.	2014
Genetic algorithm-based heuristic for feature selection in credit risk assessment	Oreski, S.; Oreski, G.	2014
Bankruptcy prediction using Extreme Learning Machine and financial expertise	Yu, Q.; Miche, Y.; Severin, E.; Lendasse, A.	2014
Combining cluster analysis with classifier ensembles to predict financial distress	Tsai, C. F.	2014
Improving experimental studies about ensembles of classifiers for bankruptcy prediction and credit scoring	Abellan, J.; Mantas, C. J.	2014
A genetic algorithm approach for SMEs bankruptcy prediction: Empirical evidence from Italy	Gordini, N.	2014

A comparative study of classifier ensembles for bankruptcy prediction	Tsai, C. F.; Hsu, Y. F.; Yen, D. C.	2014
Novel feature selection methods to financial distress prediction	Lin, F. Y.; Liang, D. R.; Yeh, C. C.; Huang, J. C.	2014
An improved boosting based on feature selection for corporate bankruptcy prediction	Wang, G.; Ma, J.; Yang, S. L.	2014
Predicting restaurant financial distress using decision tree and AdaBoosted decision tree models	Kim, S. Y.; Upneja, A.	2014
Credit risk assessment model for Jordanian commercial banks: Neural scoring approach	Bekhet, Hussain Ali; Eletter, Shorouq Fathi Kamel	2014
Development and application of consumer credit scoring models using profit-based classification measures	Verbraken, T.; Bravo, C.; Weber, R.; Baesens, B.	2014
AdaBoost based bankruptcy forecasting of Korean construction companies	Heo, J.; Yang, J. Y.	2014
Financial ratio selection for business failure prediction using soft set theory	Xu, Wei; Xiao, Zhi; Dang, Xin; Yang, Daoli; Yang, Xianglei	2014
Credit risk evaluation using multi-criteria optimization classifier with kernel, fuzzification and penalty factors	Zhang, Z. W.; Gao, G. X.; Shi, Y.	2014
Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research	Lessmann, S.; Baesens, B.; Seow, H. V.; Thomas, L. C.	2015
Credit scoring using the clustered support vector machine	Harris, T.	2015
Prediction of financial distress: An empirical study of listed Chinese companies using data mining	Geng, R. B.; Bose, I.; Chen, X.	2015
Bankruptcy visualization and prediction using neural networks: A study of U.S. commercial banks	López Iturriaga, F. J.; Sanz, I. P.	2015
Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction	Kim, M. J.; Kang, D. K.; Kim, H. B.	2015
Investigation and improvement of multi-layer perceptron neural networks for credit scoring	Zhao, Z. Y.; Xu, S. X.; Kang, B. H.; Kabir, M. M. J.; Liu, Y. L.; Wasinger, R.	2015
Variable selection and corporate bankruptcy forecasts	Tian, S. N.; Yu, Y.; Guo, H.	2015
Genetic algorithms for credit scoring: Alternative fitness function performance comparison	Kozeny, V.	2015
Selection of Support Vector Machines based classifiers for credit risk domain	Danenas, P.; Garsva, G.	2015
The effect of feature selection on financial distress prediction	Liang, D.; Tsai, C. F.; Wu, H. T.	2015
Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study	Liang, D.; Lu, C. C.; Tsai, C. F.; Shih, G. A.	2016
The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending	Serrano-Cinca, C.; Gutierrez-Nieto, B.	2016
Financial distress prediction using the hybrid associative memory with translation	Cleofas-Sánchez, L.; García, V.; Marqués, A. I.; Sánchez, J. S.	2016
The dynamic financial distress prediction method of EBW-VSTW-SVM	Sun, J.; Li, H.; Chang, P. C.; He, K. Y.	2016
A novel multistage deep belief network based extreme learning machine ensemble learning paradigm for credit risk assessment	Yu, L. A.; Yang, Z. B.; Tang, L.	2016
From footprint to evidence: An exploratory study of mining social data for credit scoring	Guo, G.; Zhu, F.; Chen, E.; Liu, Q.; Wu, L.; Guan, C.	2016
An ELM-based classification algorithm with optimal cutoff selection for credit risk assessment	Yu, L.; Li, X.; Tang, L.; Gao, L.	2016
An Effective Computational Model for Bankruptcy Prediction Using Kernel Extreme Learning Machine Approach	Zhao, D.; Huang, C.; Wei, Y.; Yu, F.; Wang, M.; Chen, H.	2017
Dynamic financial distress prediction with concept drift based on time weighting combined with Adaboost support vector machine ensemble	Sun, J.; Fujita, H.; Chen, P.; Li, H.	2017
Grey wolf optimization evolving kernel extreme learning machine: Application to bankruptcy prediction	Wang, M.; Chen, H.; Li, H.; Cai, Z.; Zhao, X.; Tong, C.; Li, J.; Xu, X.	2017

APÊNDICE II - R Script utilizado na pesquisa

```
Require (xgboost)
Library (janitor)
Library (lubridate)
Library (hms)
Library (tidyr)
Library (stringr)
Library (readr)
Library (forcats)
Library (RcppRoll)
Library (dplyr)
Library (tibble)

FINDISTR = as.logical(FINDISTR)

build_model(model_func = xgboost_binary, formula = FINDISTR ~ PMREC + ENDV + LIQIMED + CATA + COMENDIV + LIQCOR + LPA + PASPATL + PATLATÑCIR + FATA + GRSales + GRWTA + NetPCAS + NetPEQTY + NetPFIXA + NetPTOTA + MARGL + MARKVAL, eval_metric = "auc", nrounds = 1000, max_depth = 5, min_child_weight = 1, gamma = 0, learning_rate = 0.1, test_rate = 0.3)

prediction_binary(data = "test", threshold = "accuracy_rate")

prediction_binary(data = "test", threshold = "accuracy_rate") %>% do_roc(predicted_probability, calculation_1)
```

APÊNDICE III - Matriz de correlação das variáveis.

Variáveis	PMREC	ENDV	LIQIMED	CATA	COMENDIV	LPA	LIQCOR	PASPATL	PATLATÃCIR	FATA	GRSALES	GRWTA	NETPCAS	NETPEQTY	NETPFIXA	NETPTOTA	MARGL	MARKVAL	FINDISTR
PMREC	1,00																		
ENDV	0,00	1,00																	
LIQIMED	-0,01	-0,03	1,00																
CATA	0,06	-0,05	0,16	1,00															
LIQCOR	-0,01	-0,03	0,95	0,15	1,00														
COMENDIV	0,00	-0,08	-0,02	0,10	-0,02	1,00													
LPA	0,00	-0,01	0,03	0,03	0,03	0,01	1,00												
PASPATL	0,00	0,00	-0,03	0,00	-0,03	-0,04	0,01	1,00											
PATLATÃCIR	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	1,00										
FATA	-0,03	0,04	-0,12	-0,57	-0,13	-0,27	-0,02	0,02	-0,03	1,00									
GRSALES	0,00	0,00	0,00	0,01	0,00	-0,01	0,00	0,00	0,00	0,00	1,00								
GRWTA	0,00	0,00	0,01	0,02	0,01	0,00	0,00	0,00	0,00	-0,01	0,00	1,00							
NETPCAS	0,00	-0,16	0,02	0,04	0,02	0,03	0,02	0,00	0,00	-0,04	0,00	0,00	1,00						
NETPEQTY	0,00	0,00	0,01	0,02	0,01	0,00	0,02	-0,42	0,00	-0,02	0,00	0,00	0,02	1,00					
NETPFIXA	0,00	0,00	0,01	0,00	0,01	0,00	0,00	0,00	0,96	-0,02	0,00	0,00	0,00	0,00	1,00				
NETPTOTA	0,00	-0,79	0,03	0,04	0,03	0,06	0,05	-0,01	0,00	-0,03	0,00	0,00	0,18	0,04	0,00	1,00			
MARGL	0,00	-0,34	-0,03	0,05	-0,03	0,06	0,00	0,00	0,00	-0,02	0,00	0,00	0,23	0,00	0,00	0,39	1,00		
MARKVAL	-0,01	-0,02	-0,01	-0,05	-0,01	-0,06	0,02	0,00	0,00	0,07	0,00	0,00	0,01	0,00	0,00	0,04	0,01	1,00	
FINDISTR	0,03	0,06	-0,06	-0,04	-0,06	0,07	-0,07	-0,03	-0,01	0,17	-0,01	-0,01	-0,04	0,01	-0,01	-0,08	-0,01	-0,06	1,00

Nota: CATA - Ativo Circulante/ Ativo Total; COM.ENDIV - Composição do Endividamento: Passivo Circulante/ (Passivo Circulante+Exigivel a Longo prazo); EBIT/Desp.Fin - EBIT/ Despesas Financeiras; ENDV - Endividamento: Passivo Circulante + Passivo Não Circulante/Ativo total; FATA - Ativo Permanente/ Ativo Total; GR/Sales - Receita Bruta/ Vendas; GRW.TA - Taxa de Crescimento ativo total; LIQ.COR - Ativo Circulante / Passivo Circulante; LIQ.IMED - Disponível / Passivo Circulante; LPA - Lucro Por Ação; MARG.L - Margem Liquida (%); MARK.VAL - Valor de Mercado; Net.P/C.AS - Lucro Líquido/ Ativo Total Circulante; Net.P/EQTY - Lucro Líquido/ Patrimônio Líquido; Net.P/FIX.A - Lucro Líquido/ Ativo Permanente; Net.P/TOT.A - Lucro Líquido/ Ativo Total; PAS/PAT.L - Passivo/ Patrimônio Líquido; PAT.L/AT.ÑCIR - Patrimônio Líquido/ Ativo não Circulante; PMREC - Prazo Médio de Recebimento (dias)

Fonte: Resultados da pesquisa.