



***MODELO PARA PREVISÃO DE DEFAULT EM EMPRESAS BRASILEIRAS NOS ANOS DE 2013 A 2020 ATRAVÉS DE TÉCNICAS DE MACHINE LEARNING***

***MODEL FOR DEFAULT PREDICTING IN BRAZILIAN COMPANIES FROM 2013 TO 2020 THROUGH MACHINE LEARNING TECHNIQUES***

Área temática: Finanças, Contabilidade e Controladoria

*MORETE, Paulo Eduardo Damiani  
LIMA, Fabiano Guasti  
Membro da Comunidade*

**Resumo**

O objetivo deste estudo é desenvolver um modelo de classificação de risco de crédito para empresas de capital aberto não financeiras que atuam no Brasil listadas na B3, através de uma técnica de *machine learning*, linear, para prever a probabilidade de ocorrência de eventos de *default* com as empresas selecionadas no horizonte de tempo de um ano. A técnica de classificação utilizada é a regressão logística, para a construção do modelo serão incorporadas variáveis, indicadores econômico-financeiros, elaboradas a partir das demonstrações financeiras anuais consolidadas das empresas, pois além de trazerem uma boa confiabilidade, os demonstrativos também refletem a saúde financeira das empresas. O modelo foi construído com uma amostra de 60 empresas de capital aberto não financeiras classificadas como solventes ou insolventes no período de 2013 a 2020, a validação do modelo foi feita utilizando uma curva ROC e seu respectivo valor AUC. Os resultados obtidos indicam que além de o modelo de classificação de risco desenvolvido ser capaz de prever eventos de *default* com um ano de antecedência com um nível de precisão de 85%, também possui uma discriminação excelente entre os dois grupos de empresas, portanto, o resultado indica que as demonstrações financeiras apresentam informações que são confiáveis para classificar empresas como prováveis solventes ou prováveis insolventes.

**Palavras-chave:** modelo de classificação de risco de crédito; empresas de capital aberto; *machine learning*; *default*; regressão logística; curva ROC; valor AUC; indicadores econômico-financeiros.

**Abstract**

The objective of this study is to develop a credit risk classification model for non-financial publicly traded companies that operate in Brazil listed on B3, through a linear machine learning technique, to predict the probability of occurrence of default events with selected companies within a one-year time horizon. The classification technique used is logistic regression, for the construction of the model, variables, economic and financial indicators will be incorporated,



prepared from the consolidated annual financial statements of the companies, as in addition to providing good reliability, the statements also reflect the financial health of companies. The model was built with a sample of 60 non-financial publicly traded companies classified as solvent or insolvent in the period from 2013 to 2020, the model validation was performed using a ROC curve and its respective AUC value. The results obtained indicate that in addition to the developed risk classification model being able to predict default events one year in advance with an accuracy level of 85%, it also has excellent discrimination between the two groups of companies, therefore, the result indicates that the financial statements present information that is reliable to classify companies as probable solvent or probable insolvent.

**Keywords:** credit risk rating model; public companies; machine learning; default; logistic regression; ROC curve; AUC value; economic and financial ratios.



## 1. Introdução

A obtenção de crédito por parte de qualquer entidade, empresa, organização ou indivíduo tem por objetivo atender suas respectivas necessidades de caixa imediatas, sejam elas para aquisição, desde o financiamento de um ativo essencial para o processo de produção de uma indústria até a compra de um ativo pessoal, como um carro.

No Brasil, as operações realizadas no mercado de crédito, dentro de uma política de especialização do Sistema Financeiro Nacional (SFN), são tipicamente realizadas por instituições financeiras bancárias (bancos comerciais e múltiplos), mas a atuação do mercado pode ser mais abrangente, provendo recursos a médio prazo, por meio de instituições financeiras não bancárias, aos consumidores de bens de consumo (ASSAF NETO, 2017). A solicitação de um empréstimo passa por um processo de análise de crédito. No caso de uma empresa, essa avaliação é feita com base em indicadores financeiros, que permitem à instituição financeira obter um panorama da rentabilidade, atividade, grau de alavancagem, liquidez e capacidade de geração de caixa da requerente, que são elaborados a partir dos demonstrativos financeiros da empresa.

Nesse sentido, é de extrema importância para o setor bancário que o gerenciamento e a mensuração do risco de crédito sejam feitos de forma adequada e correta, visto que, a atividade bancária envolve principalmente transações de empréstimos, títulos de renda fixa, contratos de derivativos, ou seja, transações que envolvem contratos de crédito, cuja ocorrência do *default* pode resultar em perdas não só para as instituições financeiras, mas para diversas outras partes que podem estar envolvidas indiretamente com a transação.

“O risco de crédito é o risco de incorrer em perdas devido à inadimplência dos tomadores ou à deterioração da capacidade de crédito.” (BESSIS, 2010, p. 7, tradução nossa).<sup>1</sup> A partir do trecho anterior definiremos aqui o evento da inadimplência, evento de *default*, como o não cumprimento das obrigações contratuais por uma das partes contratantes, o início de um processo legal como a falência ou, a redução do valor econômico dos ativos a níveis inferiores

---

<sup>1</sup> “Credit risk is the risk of losses due to borrowers’ defaults or deterioration of credit standing.”



a suas dívidas, caracterizando uma situação de incapacidade de liquidação das suas obrigações. Sendo a perda decorrente de um evento de *default*, a perda proporcional resultante da inadimplência.

Os processos de avaliação de crédito passaram por diversas revisões, dado o grande avanço da capacidade computacional que possibilitou mensurar e avaliar maiores quantidades de dados e informações. As instituições financeiras passaram a ter uma maior ênfase em modelos quantitativos para a tomada de decisão da concessão de crédito, como a implementação de técnicas mais sofisticadas de *machine learning* com intuito de criar modelos mais eficazes e precisos em prever e mensurar o risco dos tomadores e das carteiras de crédito e assim evitar maiores perdas nos eventos de *default*.

O objetivo deste trabalho é desenvolver um modelo de classificação de risco de crédito para empresas de capital aberto que atuam no Brasil através de uma técnica de *machine learning*, linear, para prever a probabilidade de ocorrência de eventos de *default* com as empresas selecionadas no horizonte de tempo de um ano. Será utilizada a técnica de classificação linear denominada regressão logística, e para a construção do modelo serão incorporadas variáveis (indicadores) elaboradas a partir das demonstrações financeiras anuais das empresas, pois além de trazerem uma boa confiabilidade, os demonstrativos também refletem a saúde financeira das empresas.

O trabalho busca auxiliar os gestores no processo de gerenciamento de risco de crédito e decisão acerca da concessão de crédito, a fim de apresentar resultados obtidos pelo método de regressão logística aplicada em uma amostra recente de empresas brasileiras. Portanto, o estudo visa a contribuir com as demais pesquisas existentes, buscando outros estudos que proporcionem uma abordagem mais ampla acerca da implementação de técnicas de *machine learning* na análise de risco de crédito promovendo a disseminação de novas informações no mercado financeiro.

A principal contribuição deste estudo está relacionada a aplicação da técnica de *machine learning* denominada regressão logística para a construção de um modelo de classificação de risco de crédito para prever a probabilidade de *default* de empresas brasileiras de capital aberto, através de variáveis explicativas elaboradas a partir de suas respectivas demonstrações financeiras. A escolha pela utilização da técnica de regressão logística se deu pelo fato de ser uma das técnicas mais utilizadas na hora de se avaliar crédito no mercado, devido sua capacidade de analisar dados dicotômicos, de natureza binomial. A regressão logística se



tornou importante para modelos de *credit scoring*, nos quais a variável dependente é a ocorrência de inadimplência e as variáveis independentes são seus fatores explicativos.

## 2. Revisão Bibliográfica

A condição de insolvência se caracteriza como uma adversidade que as empresas possam vir a passar, é de extrema importância que exista uma maneira para prever esse tipo de evento, com o objetivo de assegurar os interesses dos controladores das empresas e dos demais agentes de mercado. Dada a relevância desse tema e da importância de haver um controle de risco de crédito de forma adequada pelos agentes financeiros, muitos estudos surgiram com o intuito de mensurar a probabilidade de *default* associada aos diversos agentes de mercado, como empresas e clientes ou até mesmo uma transação.

A maioria dos modelos já existentes para a previsão de falência estão focados no uso de demonstrativos financeiros das empresas e utilizam técnicas lineares como a análise discriminante e a regressão logística, recentemente novas técnicas não lineares de *machine learning* vêm sendo empregadas em estudos dessa área. Um dos primeiros estudos na área de previsão de falência surgiu por volta dos anos 30 com foco em índices financeiros, através da comparação entre empresas insolventes e solventes, destaca-se o estudo de FitzPatrick (1932), que apresentou dados financeiros e contábeis para 20 pares de empresas e discutiu os indicadores contábeis como indicadores de falência.

Com a implementação de técnicas estatísticas mais avançadas, a discussão acerca dos modelos de previsão de falência ganhou certa notoriedade por volta da década de 60, época em que surgiram os primeiros estudos a desenvolver um modelo sobre a metodologia de análise discriminante, destacam-se os trabalhos de Beaver (1966) e Altman (1968).

Beaver (1966) utilizou a técnica estatística de análise discriminante para apresentar evidências empíricas de que determinados índices financeiros são mais importantes que outros para a previsão de *default*. Foram selecionadas 79 empresas concordatárias e 79 não concordatárias, todas norte-americanas emparelhadas de acordo com seus respectivos setores econômicos e tamanho dos ativos. O modelo levou em conta 6 categorias de indicadores econômico-financeiros, sendo selecionados 6 indicadores financeiros, são eles: Fluxo de Caixa sobre Dívida Total; Lucro Líquido sobre Ativo Total; Dívida Total sobre Ativo Total; Capital de Giro sobre Ativo Total; Liquidez Corrente e Intervalo sem Crédito. Como resultado, Beaver



encontrou indicadores capazes de discriminar com uma antecedência de até 5 anos do evento de concordata quais empresas são concordatárias e quais são não concordatárias, destacando o indicador de “Fluxo de Caixa sobre Dívida Total” como o melhor indicador financeiro capaz de discriminar empresas concordatárias e não concordatárias.

Altman (1968) com intuito de realizar uma análise mais rigorosa, em comparação aos modelos desenvolvidos nas décadas anteriores, selecionou um conjunto de indicadores financeiros usando a abordagem de análise discriminante multivariada para o problema de previsão de falência de corporações e teorizou que os indicadores, se analisados dentro de uma estrutura multivariada, terão uma significância estatística maior do que as demais técnicas utilizadas até o momento. O estudo levou em consideração dois grupos com mesmo número de empresas concordatárias e não concordatárias, emparelhadas de acordo com o setor econômico e o tamanho do ativo das empresas. O modelo de análise discriminante resulta em uma combinação linear dos indicadores financeiros que melhor discrimina cada um dos grupos, e as variáveis independentes finais do modelo foram: Ativo Circulante menos Passivo Circulante dividido pelo Ativo Total; Lucros Retidos dividido pelo Ativo Total; EBIT dividido pelo Ativo Total; Valor de Mercado do Patrimônio Líquidos dividido pelo Valor Contábil do Passivo, e Vendas dividido pelo Ativo Total.

O resultado foi a criação de um escore, denominado *Z - Score* de Altman com a finalidade de prever falências, porém limitado à uma amostra de empresas do setor industrial.

Ohlson (1980) publicou um dos primeiros trabalhos que utilizou a regressão logística como técnica estatística. A regressão logística além de possibilitar que os resultados sejam interpretados em termos de probabilidade, tem a vantagem de contornar diversas limitações e problemas da análise discriminante, principalmente devido ao fato de possuir hipóteses mais robustas e menos rígidas em comparação com a análise discriminante, que possui pressupostos e problemas como: Normalidade das variáveis independentes; Igualdade das matrizes de variância-covariância entre os dois grupos de interesse, no caso, os dois grupos de empresas; Resultado como um *escore* com interpretação pouco intuitiva e problemas relacionados aos critérios de emparelhamento da empresas, que costumam ser arbitrários.

O estudo levou em conta uma amostra de 105 empresas concordatárias e 2058 empresas não concordatárias para o período de 1970 a 1976, todas norte-americanas. Ohlson selecionou 9 variáveis independentes: Tamanho do Ativo; Exigível Total dividido pelo Ativo Total; Capital



de Giro sobre Ativo Total; Exigível de Curto Prazo sobre Ativo Circulante; variável *dummy* ( $= 1$  se o Exigível Total for maior que o Ativo Total, e  $= 0$  caso contrário); Lucro Líquido dividido pelo Ativo Total; Fluxo de Caixa Operacional dividido pelo Exigível Total; variável *dummy* ( $= 1$  se o Lucro Líquido foi negativo nos últimos 2 anos, e  $= 0$  caso contrário) e Variação no Lucro Líquido. Apenas as variáveis Capital de Giro sobre Ativo Total; Exigível de Curto Prazo sobre Ativo Circulante e variável *dummy* ( $= 1$  se o Lucro Líquido foi negativo nos últimos 2 anos, e  $= 0$  caso contrário) não foram estatisticamente significantes. O modelo estimado foi capaz de classificar corretamente com 96,12% de precisão as empresas que vieram a falência um ano antes do evento de *default*, 95,55% dois anos antes do evento de *default* e 92,84% três anos antes do evento de *default*.

Brito e Assaf Neto (2008) desenvolveram um modelo de risco de crédito de grandes empresas para previsão de insolvência utilizando a técnica estatística de regressão logística. A população do estudo compreende as empresas de capital aberto não financeiras listadas na Bovespa, a partir da população foi formada a amostra de empresas insolventes, contendo 30 empresas que tiveram suas ações negociadas como concordatárias no período entre 1994 e 2004, a partir da amostra de empresas insolventes foram selecionadas 30 empresas solventes, uma para cada insolvente, do mesmo setor econômico e de tamanho equivalente. Os autores testaram 25 indicadores contábeis calculados com base nos demonstrativos financeiros do penúltimo exercício anterior ao evento de *default*, dos 25 indicadores 17 foram considerados estatisticamente significantes perante o teste-t de igualdade de médias ao nível de significância de 5%, ou seja, as médias de 8 índices de solventes e insolventes não apresentaram diferenças estatisticamente significantes, logo não relevantes para o modelo. O modelo final incluiu as variáveis: Lucros retidos sobre ativo total; Endividamento financeiro; Capital de giro líquido e Saldo de tesouraria sobre vendas. O percentual de acerto acumulado foi de 88,3%, tendo sido classificadas incorretamente sete empresas da amostra (11,7%), três do grupo de solventes e quatro do grupo de insolventes.

Zhu e Li (2010) utilizaram as técnicas de análise discriminante e regressão logística no contexto de análise estatística multivariada para desenvolver modelos de risco de crédito de empresas chinesas a partir de seus dados financeiros de 2009. O estudo também propôs estudar a diferença entre os resultados obtidos e a realidade observada, além de comparar a capacidade preditiva entre as duas técnicas empregadas.



A amostra foi composta por 65 empresas classificadas como ST (*Special Treatment*) e 65 empresas não-ST, emparelhadas de acordo com os critérios de setor econômico e tamanho de ativos. Todas as empresas estão listadas nas bolsas de valores de Xangai e Shenzhen, sendo que das 130 empresas, 45 classificadas como ST e 45 como não-ST compuseram a amostra de treinamento para estimar o modelo e testar a precisão na classificação, as demais empresas formaram uma amostra utilizada para testar a precisão de previsão do modelo.

Os resultados indicaram que o modelo de análise discriminante obteve uma taxa de acerto de 74,4% para a amostra de treinamento e 70,0% para a amostra de teste, já o modelo de regressão logística obteve taxas de 87,8% e 75,0%, respectivamente. Os autores ressaltaram que devido ao tamanho limitado da amostra, possa haver um certo desvio na precisão, também existem diferenças nos índices financeiros de empresas de diferentes setores, como o estudo não considerou a diferença entre empresas de diferentes setores, portanto, é possível que exista algum erro.

Cordeiro (2020) em sua dissertação de mestrado definiu como objetivo de estudo a comparação do desempenho que diferentes técnicas de *machine learning* podem apresentar quando aplicadas para a previsão de empresas entrarem em *default* com um ano de antecedência do evento no período de 2013 a 2017. Foram utilizadas duas técnicas estatísticas lineares, a regressão logística e a análise discriminante linear, e quatro técnicas de classificação não lineares, *bagging*, *random forest*, *adaboost* e *stacking*. Foram selecionadas 14.603 empresas para fazerem parte da amostra do estudo, sendo que 1.795 empresas entraram em *default* no período analisado. Para a construção dos modelos foram utilizados indicadores econômico-financeiros como variáveis explicativas que, por sua vez, foram construídas a partir das demonstrações financeiras anuais das empresas.

A partir dos resultados obtidos, foi possível concluir que a técnica com melhor desempenho foi a de *random forest*, todavia, após transformações e tratamento no balanceamento dos dados, a técnica estatística regressão logística obteve resultados próximos em termos de performance, indicando ser uma técnica satisfatória em aplicações que ainda dependem de alta explicabilidade dos modelos.



### 3. Metodologia e Dados

#### 3.1. Amostra de Empresas

O objetivo principal deste trabalho é desenvolver um modelo de risco de crédito para empresas de capital aberto não financeiras brasileiras listadas na B3. Para a construção de um modelo de risco de crédito, é extremamente importante definir o tipo de evento de *default* que se deseja prever. Nesse estudo o evento de *default* adotado é o início de um processo legal de recuperação judicial, liquidação judicial ou falência.

O modelo de risco de crédito foi elaborado a partir de uma amostra de empresas dividida em dois grupos, solventes e insolventes. As companhias definidas como insolventes são empresas de capital aberto não financeiras listadas na B3 que, no período entre 2013 e 2020 constavam no cadastro de companhias abertas da Comissão de Valores Mobiliários (CVM) como: Em liquidação judicial; em recuperação extrajudicial; em recuperação judicial ou equivalente e falidas. Ao todo foram selecionadas 30 empresas de diferentes setores econômicos.

A partir da identificação das empresas insolventes procedeu-se à separação das empresas solventes para complementar amostra. O grupo composto pelas empresas denominadas solventes foi constituído por 30 empresas que foram emparelhadas com as empresas insolventes, de modo que, selecionou-se, para cada empresa insolvente incluída na amostra, uma empresa solvente do mesmo setor econômico e de tamanho equivalente, segundo o valor dos seus ativos. A classificação setorial adotada foi a utilizada pela plataforma Economatica.

A Tabela 1 relaciona as empresas que fazem parte da amostra, assim como seus respectivos setores econômicos, suas situações e o ano do evento de *default*.

Tabela 1 - Amostra de Empresas

(continua)

Empresas Insolventes	Setor Econômico	Situação	Data Início da Situação	Empresas Solventes
Ref. Manguinhos	Petróleo e Gás	ERJ	17/01/2013	Enauta
DEXXOS	Química	ERJ	09/04/2013	Elekeiroz
Carlos Renaux	Têxtil	F	15/07/2013	Teka
Metal. Duque	Metalurgia e Siderurgia	ERJ	14/02/2014	Aliperti
Inepar	Máquinas industriais	ERJ	15/09/2014	Metalfrío
Fibam	Metalurgia e Siderurgia	ERJ	13/11/2014	Tekno
Eneva	Energia Elétrica	ERJ	12/05/2015	Engie
Sultepa	Construção	ERJ	09/07/2015	Azevedo e Travassos



Eletrosom	Comércio	ERJ	08/09/2015	B2W Digital
Lupatech	Metalurgia e Siderurgia	ERJ	11/12/2015	Mangels Industrial
Wetzel	Veículos e Peças	ERJ	11/02/2016	Schulz

Tabela 1 - Amostra de Empresas (conclusão)

Empresas Insolventes	Setor Econômico	Situação	Data Início da Situação	Empresas Solventes
Buettner	Têxtil	F	28/04/2016	Karsten
Oi	Telecomunicações	ERJ	29/06/2016	Tim Participações
Viver	Construção	ERJ	29/09/2016	Tecnisa
MMX	Mineração	ERJ	20/12/2016	Litel Participações
Pdg Realty	Construção	ERJ	02/03/2017	Cyrela Realty
Conpel	Embalagens	ERJ	18/08/2017	Sansuy
Triunfo	Transporte e Logística	ERE	03/11/2017	Ecorodovias
Pomifrutas	Agricultura	ERJ	25/01/2018	SLC Agrícola
Eternit	Minerais não metálicos	ERJ	16/04/2018	Portobello
IGB	Eletroeletrônicos	ERJ	09/05/2018	Positivo
Digitel	Eletroeletrônicos	ERJ	29/05/2018	Itautec
Saraiva	Comércio	ERJ	27/11/2018	Magazine Luiza
Fer. Heringer	Química	ERJ	06/02/2019	Nutriplant
Futuretel	Telecomunicações	ELJ	15/02/2019	Sul 116
Newtel	Telecomunicações	ELJ	19/02/2019	Algar Telecom
Brasil Pharma	Farmacêutico e Higiene	F	10/06/2019	Dimed
Bardella	Máquinas industriais	ERJ	07/08/2019	Stara
Rodovias Do Tietê	Transporte e Logística	ERJ	19/12/2019	Triângulo do Sol
João Fortes	Construção	ERJ	11/05/2020	Moura Dubeux

Nota: ELJ = Em liquidação Judicial; ERE = Em recuperação extrajudicial; ERJ = Em recuperação judicial ou equivalente; F = Falida.

Fonte: Elaborado pelo autor.

### 3.2. Dados Financeiros

Os dados respectivos as informações financeiras das empresas selecionadas foram retiradas da plataforma Economatica. Foram selecionadas informações presentes nas demonstrações contábeis consolidadas das companhias e outros indicadores já calculados pela própria plataforma. Para as empresas insolventes, foram obtidas as três últimas demonstrações contábeis anteriores ao ano do evento e, para as empresas solventes, foram coletadas as demonstrações referentes aos mesmos exercícios utilizados nas empresas insolventes com as quais foram emparelhadas.

### 3.3. Variáveis Explicativas

As variáveis explicativas utilizadas para a construção do modelo de classificação de risco de crédito são índices econômico-financeiros utilizados com o intuito de caracterizar a situação



econômico-financeira das empresas, calculadas a partir das demonstrações financeiras consolidadas publicadas pelas instituições de capital aberto selecionadas na amostra.

Foram testados 29 índices econômico-financeiros, eles compreendem indicadores de: liquidez, rentabilidade, atividade, estrutura e de fluxo de caixa. Os índices foram calculados com base nas demonstrações contábeis do penúltimo exercício anterior ao ano do evento de *default*, visto que, a utilização dos dados do penúltimo exercício no modelo assegura um prazo de antecedência de pelo menos um ano em relação ao evento de *default*, além disso, a utilização de dados referentes ao último exercício poderiam prejudicar a qualidade do modelo, uma vez que, os índices calculados para esse período já poderiam estar refletindo a situação concordatária da empresa, caso a concordata de determinada empresa tenha sido deferida antes da publicação dos demonstrativos financeiros.

As Tabelas 2 e 3 relacionam, respectivamente, os 29 índices utilizados e a notação de suas respectivas fórmulas de cálculo.

Tabela 2 – Índices econômico-financeiros

(continua)

Código	Índice	Fórmula	Referência
X1	Ativo Circulante / Ativo Total	AC/AT	C
X2	Ativo Circulante / Lucro Líquido Consolidado	AC/LL	C
X3	Ativo Circulante / Patrimônio Líquido	AC/PL	C
X4	Caixa / Ativo	CEC/AT	C
X5	Caixa / Estoques	CEC/Estoques	C
X6	Caixa / Lucro Líquido Consolidado	CEC/LL	C
X7	Capital de Giro Líquido	Capital Giro/AT	B-AN
X8	Capital de Giro / Patrimônio Líquido	Capital Giro/PL	C
X9	EBIT / Despesas Financeiras	EBIT/DF	B-AN
X10	Endividamento de Curto Prazo	PC/AT	B-AN
X11	Endividamento Total	PT/AT	B-AN
X12	Estoques / Ativo	Estoques/AT	B-AN
X13	Estoques / Lucro Líquido Consolidado	Estoques/LL	C
X14	Fluxo de Caixa Operacional / Ativo	FCO/AT	B-AN
X15	Fluxo de Caixa Operacional / Passivo	FCO/PT	B-AN
X16	Giro do Ativo	VL/AT	B-AN
X17	Liquidez Corrente	AC/PC	B-AN
X18	Liquidez Geral	(AC+RLP)/(PC+PNC)	B-AN
X19	Liquidez Imediata	CEC/PC	B-AN
X20	Liquidez Seca	(AC-Estoques)/PC	B-AN
X21	Lucro Líquido / Ativo	LL/AT	C
X22	Lucros Retidos / Ativo	(LA+RL)/AT	B-AN
X23	Margem Operacional	EBIT/VL	B-AN
X24	Participação Capital Terceiros	PT/PL	C
X25	Patrimônio Líquido / Ativo	PL/AT	B-AN



X26	Patrimônio Líquido / Exigível Total	PL/PT	B-AN
X27	Patrimônio Líquido / Lucro Líquido Consolidado	PL/LL	E
X28	Rentabilidade dos Ativos (ROA)	(LL+PAM)/AT*100	E
Tabela 2 - Índices econômico-financeiros			(conclusão)

Código	Índice	Fórmula	Referência
X29	Retorno sobre Vendas	LL/VL	B-AN

Nota: B-AN = Brito e Assaf Neto (2008); C = Cordeiro (2020); E = Economática.

Fonte: Elaborado pelo autor.

Tabela 2 - Notação das fórmulas utilizadas

Notação	
AC	Ativo Circulante
CEC	Caixa e Equivalentes de Caixa
AT	Ativo Total
PL	Patrimônio Líquido Consolidado
LL	Lucro Líquido Consolidado
EBIT	Lucros antes de Juros e Tributos
PC	Passivo Circulante
DF	Despesas Financeiras
FCO	Fluxo de Caixa gerado pelas Operações
VL	Receita de Vendas Líquida
RLP	Realizável a Longo Prazo
PNC	Passivo não Circulante
LA	Lucros Acumulados
LR	Lucros Retidos
PT	Passivo Total
PAM	Participação de Acionistas Minoritários

Fonte: Elaborado pelo autor.

Foi realizado o teste-t para igualdade de médias para duas amostras, com o objetivo de verificar se as médias dos grupos de empresas solventes e empresas insolventes são iguais estatisticamente. A hipótese nula do teste é de que as médias para os dois grupos são iguais. Se as médias dos índices de solventes e insolventes forem iguais estatisticamente, o indicador não é relevante para o modelo de risco de crédito. Para interpretar o resultado do teste, deve-se levar em conta o p-valor obtido para cada índice, dado um nível de significância de 5%, se o p-valor obtido for maior do que 0,05, não pode ser rejeitada a hipótese nula de igualdade de médias. Rejeitar a hipótese nula significa, que, seja um índice em específico, X1 ou X2 ou qualquer outro índice, então não existe uma diferença significativa estatisticamente entre as médias daquele índice para o grupo de solventes e o grupo de insolventes.



Os resultados indicam que parar o nível de significância de 5% não pode ser rejeitada a hipótese nula para os índices X1, X5, X6, X8, X9, X10, X13, X15, X16, X17, X18, X19, X20, X24, X26 e X28, uma vez que, todos apresentaram p-valor maior ou igual a 0,05. Ou seja, os valores médios desses indicadores para os grupos de solventes e insolventes não apresentaram diferenças estatisticamente significantes, portanto, eles foram removidos do modelo de risco de crédito, já os demais índices econômico-financeiros apresentaram diferenças de médias com significância estatística.

Tabela 3 - Teste-t de igualdade de médias

Índice	Estatística t	p - valor
X1	1,4553	0,1495
X2	-2,2727	0,0267
X3	-2,6192	0,01101
X4	6,5246	1,268E-08
X5	-1,2285	0,2241
X6	-1,5051	0,1376
X7	4,0465	0,0001214
X8	1,2367	0,22
X9	1,6314	0,108
X10	-0,5577	0,5786
X11	-2,9312	0,004482
X12	5,6976	3,117E-07
X13	-1,6182	0,1109
X14	6,5508	2,526E-09
X15	1,2176	0,227
X16	-1,6053	0,1115
X17	-1,4923	0,1409
X18	-1,4982	0,1394
X19	-1,2175	0,2283
X20	-1,4004	0,1666
X21	3,8619	0,0002407
X22	3,3324	0,001457
X23	5,0829	2,384E-06
X24	-1,8319	0,07198
X25	2,9312	0,004482
X26	-1,4830	0,1434
X27	-2,0727	0,04257
X28	1,4334	0,157
X29	4,5642	2,17E-05

Fonte: Elaborado pelo autor.

### 3.4. Regressão Logística



A técnica estatística utilizada para o desenvolvimento do modelo de risco de crédito foi a regressão logística. A regressão logística é uma técnica de análise multivariada utilizada para prever ou explicar a ocorrência de determinados fenômenos quando a variável dependente for de natureza binária, ou seja, pode assumir dois resultados distintos, no caso, solvente ou insolvente. A regressão logística gera uma função matemática que permite estabelecer a probabilidade de uma observação pertencer a um grupo pré-determinado em razão do comportamento de um conjunto de variáveis independentes, que são os índices econômico-financeiros.

Diferentemente de outras técnicas de regressão linear, como a análise discriminante linear, a regressão logística possui maior flexibilidade em seus pressupostos. Segundo Ohlson (1980), além da regressão logística possibilitar que os resultados sejam interpretados em termos de probabilidade, a regressão logística também permite contornar diversos pressupostos da análise discriminante, dado ao fato de possuir hipóteses mais robustas e menos rígidas em comparação à análise discriminante linear, que assume hipóteses como: normalidade das variáveis independentes e Igualdade das matrizes de variância-covariância entre os dois grupos de interesse. Ohlson (1980) ainda comenta que o resultado da equação da análise discriminante é um *score* com interpretação pouco intuitiva, visto que, é basicamente um mecanismo discriminatório ordinal.

Dentre as premissas da regressão logística, temos: relação linear entre o vetor das variáveis explicativas X e a variável dependente Y; valor esperado dos resíduos é igual a zero; ausência de autocorrelação; ausência de correlação entre os resíduos e as variáveis explicativas e ausência de multicolinearidade.

Na regressão logística, a variável dependente passa por um processo de transformação logística, em que, primeiramente, é convertida em uma razão de probabilidade (entre 0 e 1) e em seguida em uma variável de base logarítmica. Como trata-se de uma transformação não linear, os coeficientes da regressão são estimados pelo método de máxima verossimilhança. O modelo de regressão logística é dado por:

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1X_1 + b_2X_2 + \dots + b_kX_k \quad (1)$$



Onde:  $p$  é a probabilidade de ocorrer o evento;  $1 - p$  é a probabilidade de o evento não ocorrer;  $\left(\frac{p}{1-p}\right)$  é a razão de probabilidades;  $X_i$  as variáveis independentes e  $b_i$  são os coeficientes estimados.

Os coeficientes medem o efeito de variações nas variáveis independentes sobre o logaritmo natural da razão de probabilidades, denominado *logit*. Para avaliar o impacto dos parâmetros sobre a probabilidade de ocorrer o evento, eles devem ser transformados por meio de antilogaritmo. A probabilidade associada à ocorrência do evento de interesse, no caso, de pertencer a amostra concordatória, é dada por:

$$p = \frac{1}{1 + e^{-(b_0 + b_1X_1 + b_2X_2 + \dots + b_kX_k)}} \quad (2)$$

Onde:  $e$  é a base dos logaritmos naturais (aproximadamente 2,718).

### 3.5. Modelo

Após a definição da amostra de empresas e das variáveis independentes, foram realizados diferentes testes, através da utilização do *software* de programação R, com o objetivo de seguir precisamente às premissas do modelo de regressão logística.

No modelo de regressão logística, a variável dependente pode assumir um valor entre zero e um, no caso, foi designado o número zero para o grupo de empresas insolventes, e o número um para o grupo de empresas solventes. Portanto, o ponto de corte do modelo é 0,5, empresas com resultado superior a 0,5 são classificadas como solventes e inferior a 0,5 classificadas como insolventes.

Tabela 5 - Ponto de Corte

Resultado	Classificação
> 0,5	Solvente
< 0,5	Insolvente

Fonte: Elaborado pelo autor.

Após retirar as variáveis cujos resultados do teste-t indicaram que, ao nível de significância de 5%, não pode ser rejeitada a hipótese nula de igualdade de médias, deu-se início ao processo de verificação de uma importante premissa do modelo, a ausência de multicolinearidade.

Foi utilizada a função *alias()* que tem como finalidade encontrar termos linearmente dependentes em um modelo linear especificado por uma fórmula. Os resultados mostraram que



foi encontrada dependência da variável X25 com as demais variáveis e, portanto, ela foi retirada do modelo.

Depois de retirar a variável X25, foi realizado o teste VIF (*Variance Inflation Factor*) que mede o quanto o comportamento (variância) de uma variável independente é influenciado, ou inflado, por sua interação/correlação com outras variáveis independentes. Na existência de multicolinearidade, o VIF será elevado para as variáveis independentes envolvidas, após a identificação das variáveis altamente correlacionadas, várias abordagens podem ser usadas para resolver o problema da multicolinearidade, nesse estudo, optou-se por excluir as variáveis que apresentaram VIF muito elevado.

O critério utilizado para a exclusão das variáveis seguiu uma determinada ordem. As variáveis que apresentarem os valores mais elevados no teste seriam selecionadas como possíveis candidatas à remoção. Comparando apenas as variáveis infladas, a variável identificada, do ponto de vista teórico, como similar as demais variáveis deveria ser excluída, seguindo a hipótese de que, como ela é similar a outras variáveis, provavelmente pode ser identificada como uma variável redundante. Em seguida um novo teste será realizado para identificar novas candidatas à exclusão, até que se chegue ao ponto em que nenhuma variável independente apresente um valor para o teste VIF maior ou igual a 4.

Tabela 6 - Primeiro teste VIF

Índice	VIF
X2	19,351
X3	2,202
X4	2,563
X7	4,610
X11	4,952
X12	2,544
X14	3,121
X21	10,307
X22	2,905
X23	6,423
X27	19,217
X29	7,492

Fonte: Elaborado pelo autor.

Os valores obtidos para X2, X21 e X27 foram extremamente altos, decidiu-se retirar a variável X2, visto que, levando em conta o modelo teórico, ela seria uma variável um pouco similar a outras variáveis presentes no modelo.



Tabela 7 - Segundo Teste VIF

(continua)

<b>Índice</b>	<b>VIF</b>
X3	2,130
X4	3,045
X7	3,576
X11	4,629
X12	2,449

Tabela 7 - Segundo Teste VIF

(conclusão)

<b>Índice</b>	<b>VIF</b>
X14	3,168
X21	10,946
X22	3,342
X23	6,403
X27	1,410
X29	8,216

Fonte: Elaborado pelo autor.

Mesmo após retirar o índice X2 o valor obtido para o VIF do índice X21 continuou alto, portanto, ele também foi removido.

Tabela 8 - Terceiro Teste VIF

<b>Índice</b>	<b>VIF</b>
X3	1,361
X4	3,121
X7	3,210
X11	3,078
X12	1,650
X14	2,618
X22	3,588
X23	6,780
X27	1,223
X29	7,967

Fonte: Elaborado pelo autor.

Por fim, a variável X29 também foi retirada, devido ao resultado obtido no teste VIF ainda ser elevado.

Tabela 4 - Quarto Teste VIF

(continua)

<b>Índice</b>	<b>VIF</b>
X3	1,317
X4	2,681



X7	1,625
X11	2,519
X12	1,594
X14	1,962
X22	3,749
X23	1,562
X27	1,088

Fonte: Elaborado pelo autor.

## 4. Estimação e Resultados

### 4.1. Resultados da Regressão

O modelo final foi composto pelo intercepto e 9 variáveis explicativas, sendo elas: X3 (Ativo Circulante/Patrimônio Líquido), X4 (Caixa/Ativo Total), X7 (Capital de Giro Líquido), X11 (Endividamento Total), X12 (Estoques/Ativo Total), X14 (Fluxo de Caixa Operacional/Ativo Total), X22 (Lucros Retidos/Ativo Total), X23 (Margem Operacional) e X27 (Patrimônio Líquido/Lucro Líquido Consolidado). A função matemática do modelo de regressão logística é dada por:

$$\ln\left(\frac{p}{1-p}\right) = -1,675 - 0,228X_3 + 32,302X_4 + 0,272X_7 + 1,022X_{11} + 2,938X_{12} - 0,642X_{14} + 1,439X_{22} - 0,144X_{23} - 0,006X_{27} \quad (3)$$

As variáveis X3, X14, X23 e X27 apresentaram coeficientes negativos, o que significa que, quanto maiores forem os valores desses índices, menor será a probabilidade da ocorrência do evento de *default*, no entanto, as variáveis X4, X7, X11, X12 e X22 apresentaram coeficientes com valores positivos, logo, quanto maiores forem os valores desses índices, maior será a probabilidade de a empresa sofrer um *default*.

### 4.2. Matriz de Confusão

Com o objetivo de avaliar a capacidade de previsão do modelo, foi construída uma matriz de confusão.

Tabela 5 - Matriz de Confusão

Observado	Predito		Total	Precisão
	Insolvente	Solvente		
Insolvente	27	3	30	90,00%
Solvente	6	24	30	80,00%
Total	33	27	60	85,00%



Percentual de Acerto do Modelo	85,00%
95% CI	(0,7343 , 0,929)
<i>P-Value [Acc &gt; NIR]</i>	1,54E-05
Sensitividade	90,00%
Especificidade	80,00%

Fonte: Elaborado pelo autor.

Os resultados indicam que o modelo tem uma taxa de acerto de 85%, classificando corretamente 51 das 60 empresas da amostra. Do grupo de empresas insolventes, foram classificadas corretamente 27 das 30 empresas da amostra, o que representa uma sensibilidade de 90%, ou seja, uma taxa de acerto de 90% no evento de interesse, que, no caso, é classificar a empresa como insolvente, já do grupo de empresas solventes, foram classificadas corretamente 24 das 30 empresas da amostra, especificidade de 80%, ou seja, uma taxa de acerto de 80% em classificar a empresa como solvente. Para um Intervalo de Confiança de 95% o modelo tem uma taxa de assertividade que varia entre 73,43% e 92,90%. O p-valor foi de 1,543e-08, o que indica que o modelo é relevante estatisticamente.

Quando comparados os resultados obtidos neste trabalho com os resultados dos estudos mencionados anteriormente que, ao menos, utilizaram amostras de empresas emparelhadas entre dois grupos, de acordo com o setor econômico, tamanho de seus ativos e utilizaram uma técnica estatística, que, seja uma técnica de *machine learning* linear ou a técnica de regressão logística. É possível afirmar que os resultados obtidos são satisfatórios.

Brito e Assaf Neto (2008) selecionaram 30 empresas concordatárias e 30 não concordatárias, brasileiras e de capital aberto, emparelhadas segundo o critério de setor econômico e tamanho dos ativos e obtiveram um nível de acerto de 88,3%. Zhu e Li (2010) escolheram empresas listadas nas bolsas de Xangai e Shenzhen emparelhadas de acordo com o setor e tamanho de ativos, e obtiveram um nível de acurácia de 87,8% 75,0% para a amostra de treinamento e amostra de teste, respectivamente. Os resultados desses dois estudos não ficaram muito distantes do resultado obtido neste trabalho, que apresentou um nível de acerto de 85%.

### 4.3. Medidas de Ajuste

Embora não haja um consenso na literatura sobre como avaliar o ajuste de uma regressão logística, existem algumas abordagens. A qualidade do ajuste do modelo de regressão logística pode ser expressa por algumas variantes de uma medida de ajuste denominada como: Pseudo



$R^2$  da estatística, que são medidas que se assemelham ao coeficiente de determinação de uma regressão linear, a maioria dessas variantes é baseada no desvio do modelo. Além da medida dos Pseudo  $R^2$ , a avaliação do nível de ajuste de um modelo de regressão logística pode ser realizada por meio do teste de Hosmer e Lemeshow (1980).

Para o modelo de regressão logística em questão foram calculados os Pseudo  $R^2$  de McFadden, Cox-Snell e Nagelkerke, também foi realizado o teste de Hosmer e Lemeshow.

Tabela 11 - Pseudo  $R^2$

Pseudo $R^2$	
McFadden $R^2$	0,352896
Cox-Snell $R^2$	0,386895
Nagelkerke $R^2$	0,515859

Fonte: Elaborado pelo autor.

Tabela 12 - Teste Hosmer e Lemeshow

Teste de Hosmer e Lemeshow		
Qui - Quadrado	g.l.	P - Valor
14,562	8	0,06823

Nota: g.l. = Graus de liberdade.

Fonte: Elaborado pelo autor.

O Cox-Snell  $R^2$  baseia-se no *Likelihood Value* para o modelo em comparação com o *Likelihood Value* para um modelo base, no entanto, com resultados categóricos. Esse Pseudo  $R^2$  tem uma escala que começa em zero, porém, seu valor máximo teórico é inferior a um, mesmo para um modelo que possua um ajuste dito como perfeito. O Cox-Snell  $R^2$  trata-se de uma medida que pode ser utilizada para comparar o desempenho de modelos concorrentes, sendo preferido aquele que apresente o Cox-Snell  $R^2$  mais elevado. A interpretação que temos a partir dos resultados obtidos é a de que aproximadamente 38,89% das variações ocorridas são explicadas pelo conjunto das variáveis independentes.

O  $R^2$  de Nagelkerke é uma versão ajustada do Cox-Snell  $R^2$  que ajusta a escala da estatística para cobrir um intervalo de zero a um. A interpretação obtida a partir dos resultados é a de que o modelo é capaz de explicar aproximadamente 51,59% das variações registradas na variável dependente.

O  $R^2$  de McFadden é uma outra versão baseada nos *kernels* de máxima verossimilhança para o modelo somente de interceptação e o modelo estimado completo. A interpretação dos resultados



é a de que o modelo é capaz de explicar aproximadamente 35,29% das variações registradas na variável dependente.

Vale ressaltar que para todos os Pseudo  $R^2$ , maiores valores dessas medidas indicam um melhor ajuste do modelo.

O teste de Hosmer e Lemeshow (1980) mede o grau de acurácia do modelo de regressão logística, este indicador corresponde a um teste Qui-Quadrado que consiste em dividir o número de observações em classes e, em seguida, comparar as frequências previstas com as observadas. A finalidade desse teste é verificar se existem diferenças significativas entre as classificações realizadas pelo modelo e a realidade observada. A hipótese nula do teste é a de que não há diferenças significativas entre os resultados previstos pelo modelo e os observados, portanto, busca-se não rejeitar a hipótese de que não existem diferenças entre os valores previstos e observados.

Como o p-valor obtido é superior a 0,05, ao nível de significância de 5%, aceita-se a hipótese nula de que não há diferenças significativas entre os valores previstos e os valores observados, portanto, estatisticamente, não existem diferenças significativas entre os valores previstos e os valores observados, pode-se dizer que o modelo é capaz de produzir estimativas e classificações confiáveis.

#### **4.4. Curva ROC e valor AUC**

A técnica de ranqueamento escolhida para avaliar a performance do modelo foi a curva ROC (*Receiver Operating Characteristic*) e sua área AUC (*Area under the curve*).

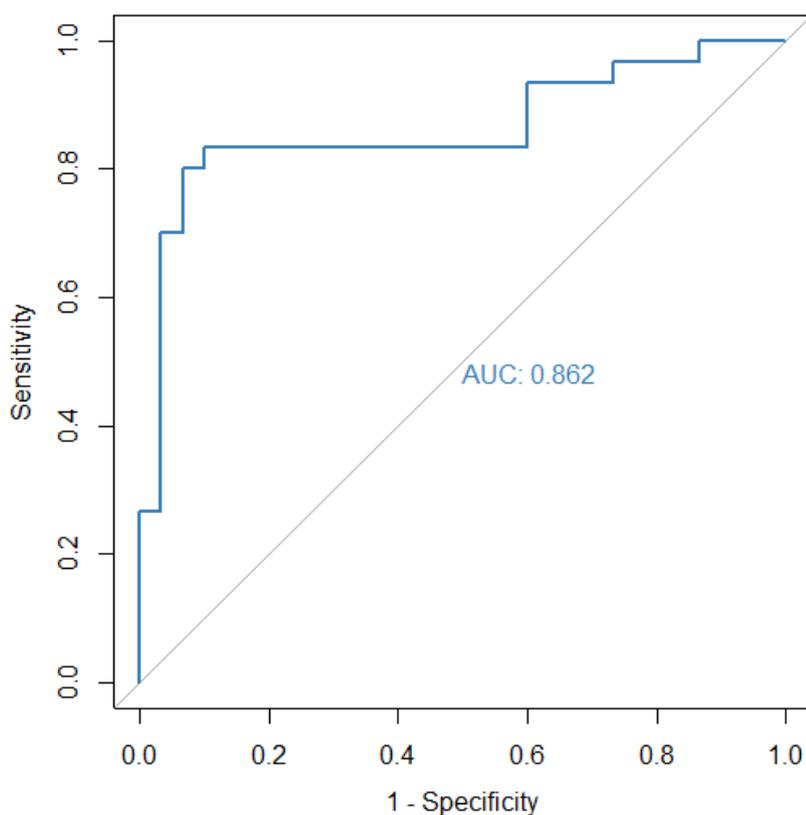
A curva ROC é um gráfico que resume o desempenho de um modelo de classificação binário em relação à classe positiva (previsão, pelo modelo, de ocorrência do evento de interesse, no caso, classificação da empresa como insolvente). O eixo x indica a especificidade que é a proporção de acerto na previsão da não ocorrência de um evento nos casos em que ele de fato não ocorreu e o eixo y indica a sensibilidade, que é a proporção de acerto na previsão da ocorrência de um evento nos casos em que ele de fato ocorreu.

Podemos interpretar a curva ROC como a proporção de previsões corretas para a classe positiva (eixo y) *versus* a proporção de erros da classe negativa (eixo x). Portanto, o melhor modelo possível é aquele que alcança o canto superior esquerdo do gráfico, ou seja, o modelo que é capaz de classificar perfeitamente todos os casos. Portanto, quanto maior a área abaixo da Curva



ROC, ou seja, quanto maior a área AUC, maior é a capacidade do modelo em discriminar os grupos de evento de interesse e não interesse.

Figura 1 - Curva ROC e valor AUC



Fonte: Elaborado pelo autor.

Tabela 13 - Interpretação do valor AUC

Área abaixo da curva ROC	Interpretação
0,5 a 0,6	E: Não há discriminação
0,6 a 0,7	D: Discriminação pobre
0,7 a 0,8	C: Aceitável/Regular
0,8 a 0,9	B: Excelente/Bom
0,8 a 0,9	A: Excepcional

Fonte: Elaborado pelo autor com base em LANTZ (2019).

O valor abaixo da curva ROC do modelo foi de 0,862, o que indica que existe uma excelente discriminação entre os dois grupos pelo modelo.



Ao comparar os resultados obtidos neste trabalho da curva ROC e seu respectivo valor AUC, com os resultados dos estudos anteriores mencionados nessa monografia que utilizaram a curva ROC e seu respectivo valor AUC, vemos que, apesar dos valores diferirem bastante, no mínimo todos possuem uma discriminação aceitável.

Brito e Assaf Neto (2008), que disponibilizaram da mesma técnica estatística e utilizaram métricas para a seleção da amostra de empresas similares que a desta monografia, obtiveram um valor AUC de 0,978, o que indica que o modelo desenvolvido possui uma discriminação excepcional. Cordeiro (2020) disponibilizou de duas técnicas estatísticas lineares, a regressão logística e a análise discriminante linear, e quatro técnicas de classificação não lineares, *bagging*, *random forest*, *adaboost* e *stacking*, e obteve os seguintes resultados para o valor da área abaixo da curva ROC: 0,750 (Regressão logística), 0,750 (Análise discriminante linear), 0,761 (*bagging*), 0,769 (*random forest*), 0,731 (*adaboost*) e 0,746 (*stacking*). Apesar dos valores obtidos serem inferiores ao desta monografia, vale ressaltar que Cordeiro (2020) selecionou uma amostra de 14.603 empresas, com muitas delas apresentando dados referentes as demonstrações financeiras faltantes ou desbalanceados, sendo que em muitos casos não existe uma empresa de auditoria por trás da revisão dos dados contábeis das empresas.



## 5. Conclusão

O objetivo deste estudo é desenvolver um modelo de classificação de risco de crédito para empresas de capital aberto não financeiras que atuam no Brasil listadas na B3, utilizando indicadores financeiros calculados a partir das demonstrações financeiras anuais como variáveis explicativas, para prever a probabilidade de ocorrência de eventos de default com as empresas selecionadas no horizonte de tempo de um ano. A técnica de classificação utilizada é a Regressão Logística.

As variáveis explicativas são: X3 (Ativo Circulante / Patrimônio Líquido), X4 (Caixa / Ativo Total), X7 (Capital de Giro Líquido), X11 (Endividamento Total), X12 (Estoques / Ativo Total), X14 (Fluxo de Caixa Operacional / Ativo Total), X22 (Lucros Retidos / Ativo Total), X23 (Margem Operacional) e X27 (Patrimônio Líquido / Lucro Líquido Consolidado).

É possível concluir, a partir dos resultados obtidos, que a utilização de indicadores contábeis é uma boa abordagem para a construção de modelos de classificação de risco de crédito, uma vez que, os resultados indicam que essa abordagem permitiu classificar com um bom nível de precisão empresas como solventes e insolventes, para a amostra utilizada nesse estudo o nível de acerto do modelo foi de 85%.

Apesar de os resultados indicarem que a construção de um modelo de risco de crédito a partir de indicadores financeiros tenha um elevado grau de precisão e, portanto, uma excelente discriminação, é importante ressaltar que todas as empresas selecionadas para a amostra são empresas não financeiras e de capital aberto que estão listadas na B3, ou seja, existem empresas de auditoria que asseguram um alto grau de qualidade e confiabilidades das informações financeiras disponibilizadas publicamente, o que não garante que este modelo é adequado para empresas não auditadas, dado que seus indicadores contábeis podem não representar a verdadeira situação financeira da empresa.



## Referencial Bibliográfico

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

ASSAF NETO, A. Mercados Financeiros: Monetário e Crédito. *in*: ASSAF NETO, A. **Mercado Financeiro**. 13. ed. São Paulo: Atlas, 2017. cap. 4, p. 66-86.

BEAVER, W. H. Financial Ratios as Predictors of Failure. **Journal of Accounting Research**, Wiley on behalf of Accounting Research Center, Booth School of Business, University of Chicago, v. 4, p. 71-111, 1966.

BESSIS, J. The 2007–2008 Financial Crisis. *in*: BESSIS, J. **Risk Management in Banking**. 3. ed. Chichester: John Wiley & Sons, 2010. cap. 1, p. 3-18.

BRITO, G. A. S.; ASSAF NETO, A. Modelo de classificação de risco de crédito de empresas. **Revista Contabilidade & Finanças**, São Paulo, v. 19, n. 46, p. 18-29, jan./abr. 2008.

CORDEIRO, T. V. B. **Predição de Default de Empresas: Técnicas de Machine Learning em Dados Desbalanceados**. 2020. 65 p. Dissertação (Mestrado Profissional em Finanças e Economia) - Escola de Economia de São Paulo, Fundação Getúlio Vargas, São Paulo, 2020.

FITZPATRICK, P. J. A comparison of the ratios of successful industrial enterprises with those of failed companies. **Certified Public Accountant**, v. 2, p. 598–605, 1932.

HOSMER, D. W.; LEMESHOW, S. A Goodness-of-Fit Test for the Multiple Logistic Regression Model. **Communications in Statistics**, v. 10, p. 1043-1069, 1980.



LANTZ, B. Evaluating Model Performance. *in*: LANTZ, B. **Machine Learning with R: Expert techniques for predictive modeling**. 3 ed. Birmingham: Packt Publishing, 2019. cap. 10, p. 313-346.

OHLSON, J. A. Financial Ratios and the Probabilistic Predictions of Bankruptcy. **Journal of Accounting Research**, Wiley on behalf of Accounting Research Center, Booth School of Business, University of Chicago, v. 18, n. 1, p. 109-131, 1980.

ZHU, K. L.; LI, J. J. Studies of Discriminant Analysis and Logistic Regression Model Application in Credit Risk for China's Listed Companies. **Management Science and Engineering**, v. 4, n. 4, p. 24-32, jul./out. 2010.