*Revista Gestão & Tecnologia*
*e-ISSN: 2177-6652*
*revistagt@fpl.edu.br*
*http://revistagt.fpl.edu.br/*

# ANALISANDO MÉTODOS DE *MACHINE LEARNING* E AVALIAÇÃO DO RISCO DE CRÉDITO

# ANALYSING MACHINE LEARNING METHODS AND CREDIT RISK ASSESSMENT

# ANALISANDO MÉTODOS DE MACHINE LEARNING Y EVALUACIÓN DEL RIESGO DE CRÉDITO

Felipe Fernandes Coelho
Pós-graduando em Gestão com ênfase em Finanças na Fundação Dom Cabral
felipefcoelho10@gmail.com

Daniel Penido de Lima Amorim
Economista, mestre em Administração / Finanças pela Universidade Federal de Minas Gerais (UFMG).
daniel_amorim23@hotmail.com
http://orcid.org/0000-0002-2844-3079

Marcos Antônio de Camargos
Doutor e Mestre em Administração pelo CEPEAD-UFMG. Coordenador do Mestrado Profissional em Economia e Professor Titular da Faculdade IBMEC-MG.
marcosac@face.ufmg.br
https://orcid.org/0000-0002-3456-8249

Editor Científico: José Edson Lara
Organização Comitê Científico
Double Blind Review pelo SEER/OJS
Recebido em 23.01.2021
Aprovado em 02.03.2021

**Resumo**

**Objetivo do estudo:** O objetivo deste artigo é comparar a regressão logística clássica e dois métodos de *machine learning* para *credit scoring*, o *random forest* e o *XGBoost*, visando identificar qual apresenta melhor desempenho na previsão de inadimplência.

**Metodologia/abordagem:** O desempenho dos modelos estimados foi comparado com base em acurácia, estatística Kolmogorov-Smirnov, além de curva ROC.

**Originalidade/Relevância:** Foi utilizada uma base de dados exclusiva com informações de 3.844 pequenas e médias empresas, clientes de uma locadora de automóveis com atuação em todo o Brasil.

**Principais resultados:** Os resultados sugerem que os métodos de *machine learning* apresentam capacidade preditiva maior quando comparados com a regressão logística. O *XGBoost* teve o melhor desempenho, entre os métodos analisados.

**Contribuições teóricas/metodológicas:** Este artigo corrobora a utilização de variáveis não financeiras para a previsão de inadimplência e a superioridade dos métodos estatísticos mais modernos frente à abordagem clássica.

**Palavras-chave:** Risco de Crédito; *Credit Scoring*; Regressão Logística; *Machine Learning; XGBoost*; *Random Forest*.

**Abstract**

**Objective:** In this article, we compare the classic logistic regression and two machine learning methods for a credit scoring model, random forest, and XGBoost, aiming to identify which one presents the best performance in predicting defaults.

**Methodology/approach:** We compared the performance of the estimated models based on accuracy, and Kolmogorov-Smirnov statistic, besides the ROC curve.

**Originality/Relevance:** We use an exclusive database with information of 3,844 small and medium-sized companies, clients of a car rental company that operates in the whole of Brazil.

**Main results:** The results suggested that machine learning methods have a greater predictive capacity when compared to logistic regression. XGBoost had the best performance among the analyzed methods.

**Theoretical contributions:** This article corroborates the utilization of non-financial variables for default prediction and the superiority of the most modern statistical methods compared to the classical approach.

**Keywords:** Credit Risk; Credit Scoring; Logistic Regression; Machine Learning; XGBoost; Random Forest.

**Resumen**
**Objetivo del estudio:** El objetivo de este artículo es comparar la Regresión Logística clásica y dos métodos de *machine learning* para el *scoring* de crédito, *random forest* y *XGBoost*, con el fin de identificar cuál tiene el mejor desempeño en la en la predicción de impagos.

**Metodología/enfoque:** El desempeño de los modelos estimados se comparó en base a la exactitud, la estadística Kolmogorov-Smirnov y la curva ROC.

**Originalidad/Relevancia:** Se utilizó una base de datos exclusiva con informaciones de 3.844 pequeñas y medias empresas, clientes de una compañía de alquiler de automóviles que opera en todo Brasil.

**Resultados principales:** Los resultados sugieren que los métodos de *machine learning* tienen una capacidad de predicción mayor en comparación con la Regresión Logística. El *XGBoost* obtuvo el mejor desempeño entre los métodos analizados.

**Contribuciones teóricas/metodológicas:** Este artículo corrobora el uso de variables no financieras para la previsión de impagos y la superioridad de los métodos estadísticos más modernos frente al enfoque clásico.

**Palabras clave:** Riesgo de Crédito; S*coring* de crédito; Regresión Logística; *Machine Learning*; *XGBoost*; *Random Forest*.

## 1 INTRODUÇÃO

A crescente integração econômica e financeira em nível mundial, proporcionada pelo avanço das tecnologias da informação tem resultado em um cenário de constante mudança, abalado, em alguns momentos, por crises financeiras, econômicas e, mais recentemente, sanitárias (a exemplo, da pandemia de COVID-19). Esses eventos têm impactos diretos sobre a sociedade, especialmente, sobre a economia e o mercado financeiro, com destaque para a capacidade de pagamento de empresas e pessoas físicas.

A instabilidade que permeia um ambiente econômico de crise leva tanto pessoas físicas, quanto empresas a enfrentar dificuldades financeiras e, numa situação mais extrema, elas podem chegar à falência. Esse contexto negativo resulta no desestímulo das vendas via crédito pelas empresas, bem como no aumento do risco e do custo do crédito no mercado financeiro. Por consequência, esse cenário exige o aprimoramento de modelos cada vez mais eficientes para a análise e concessão do crédito.

O mundo dos negócios, em especial, o segmento financeiro, tem experimentado um expressivo crescimento da quantidade de informações e do aprimoramento de técnicas cada

vez mais assertivas na avaliação e recuperação do crédito. Nesse contexto, avaliar o risco de inadimplência, não somente mediante os critérios convencionais, mas também mediante a utilização de métodos estatísticos sofisticados, pode contribuir para a solidez de instituições financeiras e das empresas em geral. As abordagens de análise que utilizam *machine learning* têm sido cada vez mais empregadas para avaliar o perfil dos clientes, principalmente, como uma resposta aos ambientes dinâmicos, nos quais decisões devem ser tomadas rapidamente.

A ideia por trás do *machine learning* é o desenvolvimento de algoritmos que melhoram automaticamente com a experiência. Trata-se de uma área do conhecimento que se situa na interseção da ciência da computação e estatística, no centro da inteligência artificial e da ciência de dados, que está em rápida expansão na atualidade (Jordan & Mitchell, 2015).

O objetivo deste artigo é comparar a regressão logística clássica e os métodos de *machine learning* para *credit scoring*, identificando qual apresenta melhor desempenho ao estimar a probabilidade de inadimplência com base em dados de uma amostra de clientes pessoas jurídicas. Quanto aos métodos de *machine learning*, avaliou-se os algoritmos *Extreme Gradient Boost* (*XGBoost*) e *random forest*. Três medidas de ajuste foram adotadas na comparação do desempenho dos modelos: acurácia, estatística Kolmogorov-Smirnov, além de curva ROC. Foi analisada uma amostra composta por micro, pequenas e médias empresas (MPMEs) que dispunham de limite de crédito de até R$ 25 mil junto a uma empresa operadora do segmento de locação de automóveis, atuante em todo o território brasileiro.

Os resultados encontrados neste estudo evidenciaram que os métodos de *machine learning* apresentaram melhor desempenho que a análise convencional, baseada na regressão logística, especialmente, no que diz respeito ao *GXBoost*. Nesse sentido, este estudo contribui com um direcionamento sobre o método que pode prover resultados melhores na análise de risco de crédito.

Após essa introdução, a segunda seção do artigo apresenta uma revisão de literatura sobre crédito, *credit scoring* e aplicação de métodos estatísticos na avaliação do risco de crédito; a terceira seção apresenta os métodos adotados neste estudo; a quarta seção apresenta os dados e variáveis; a quinta apresenta e discute os resultados encontrados; e, por fim, a sexta seção tece as considerações finais.

## 2 REVISÃO DA LITERATURA

### 2.1 Crédito, *credit scoring e machine learning*

O risco de crédito é definido como o risco de perda resultante de falha entre alguma das contrapartes no cumprimento das obrigações (Qu, 2008). Financeiras, agências de classificação de risco e mesmo algumas empresas não financeiras, adotam *ratings* no intuito de indicar a probabilidade de um cliente se tornar inadimplente. Quanto maior o risco de inadimplência de um cliente, pior é seu *rating*. A deterioração do *rating* de crédito do tomador não resulta em uma perda imediata. Contudo, ela sugere um incremento na probabilidade de que a inadimplência venha a ocorrer. Cada empresa utiliza sua própria classificação de inadimplência, comumente, associada ao atraso no pagamento de um compromisso assumido pelo tomador, seja por períodos de 60 ou 90 dias (Brito & Assaf Neto, 2008).

Na concessão de crédito, o tomador pode não honrar com suas obrigações, seja por má-fé ou dificuldades financeiras, entre outros motivos. Além disso, pode ser difícil para a empresa recuperar o crédito de um cliente inadimplente, e o não pagamento de uma dívida implica prejuízos ao credor. Para a concessão de crédito ser assertiva é necessária uma análise quantitativa, que deve ser realizada rapidamente, usufruindo da quantidade e da qualidade das informações disponíveis. Esse tipo de análise permite identificar padrões que auxiliam na previsão da inadimplência, contribuindo, portanto, para a segurança e assertividade na decisão do credor.

A concessão de crédito envolve tomada de decisões complexas e dinâmicas, seja por instituições financeiras, seja por empresas a seus clientes. Essa atividade tem exigido modelos de *credit scoring* cada vez mais eficientes.

Lewis (1992) define *credit scoring* como um processo em que informações sobre o solicitante são convertidas em uma pontuação (*score*) que mostra o perfil de risco do solicitante do crédito. Os modelos de *credit scoring* visam fornecer a probabilidade de inadimplência de clientes, visando minimizar a possibilidade de perda. Portanto, tais modelos desempenham um papel crucial para a sustentabilidade de instituições financeiras e de empresas.

Segundo Hand e Adms (2000), o *credit scoring* consiste no uso de métodos estatísticos para classificar requerentes de crédito em categorias de *bons* ou *maus* tomadores, sendo uma ferramenta essencial para a tomada de decisão. Porém, Souza e Chaia (2000) ressaltam que,

apesar de o *credit scoring* representar um processo estatístico, ele não inibe a possibilidade de se recusar um *bom* pagador ou se aceitar um *mau* pagador.

Na evolução dos modelos de *credit scoring*, pode-se dizer que os primeiros trabalhos dedicados a estudar e analisar a capacidade preditiva de métodos para segregar uma amostra de dados em grupos foram realizados por Fisher (1936), que tentando classificar variedades de plantas, lançou as bases da análise discriminante e Durand (1941) que posteriormente aplicou esta nova técnica na concessão de empréstimos no mercado financeiro. Também de forma pioneira, já na década de 1960, Altman (1968) conferiria avanços no desenvolvimento de modelos de *credit scoring* utilizando a análise discriminante para a previsão de falências de grandes empresas. A partir da década de 70, passou a predominar o uso da regressão logística e da análise discriminante. A década de 90 foi marcada pelo surgimento e disseminação do uso das redes neurais. Recentemente, têm se destacado as tecnologias ligadas à inteligência artificial, como o *machine learning*, que visam melhorar a capacidade preditiva dos modelos estatísticos (Altman & Saunders, 1998; Zhong et al., 2014).

Desde então, tem sido constante o desenvolvimento e aprimoramento de técnicas e modelos que buscam classificar corretamente bons e maus pagadores na concessão de crédito. O aprimoramento dos modelos de *credit scoring*, com a intensificação do uso de ferramentas matemáticas e estatísticas, tem sido fundamental para uma concessão de crédito mais assertiva. Concernente a isso, Tsai et al. (2014) argumentam que pequenas melhorias na precisão da classificação de crédito podem resultar em grande redução do risco e de perdas para empresas e instituições financeiras.

Até pouco tempo atrás, o desenvolvimento e calibragem dos modelos de *credit scoring* ficavam a cargo de especialistas, consumindo muitos recursos e tempo. Mas, atualmente, o surgimento de algoritmos como o *machine learning* e a inteligência artificial tem contribuído para o avanço desse campo, ajudando os especialistas e reduzindo a demanda de tempo e trabalho (Munkhdalai et al., 2019).

A capacidade de utilização de dados e estatística para prever a inadimplência, contando com a sofisticação dos algoritmos de *machine learning*, faz com que a concessão de crédito se torne muito mais segura e diminua significativamente a inadimplência de instituições financeiras e de demais empresas que dispõem de um bom sistema de avaliação de crédito. Isso também favorece um melhor gerenciamento da carteira de clientes, assim como o processo de cobrança.

Conforme destaca Mitchell (1997), os algoritmos de *machine learning* constituem um sub-campo da inteligência artificial, que ao contrário das técnicas estatísticas tradicionais, não exigem o conhecimento das relações entre as variáveis de entrada e de saída dos modelos (Aniceto, 2016).

Guégan e Hassani (2018) destacam que o *machine learning* e a inteligência artificial têm a automação por traz dos seus algoritmos, cujo objetivo é aprender com os dados e fazer previsões a partir deles, imitando os processos cognitivos realizados pelo cérebro humano. Para isso, operam de forma dinâmica, adaptando-se às mudanças nos dados, não só contando com estatísticas, mas também com otimização matemática. Dessa forma, seu uso nos modelos de *credit scoring* pode potencializar sua assertividade e eficiência, principalmente, na concessão de crédito para MPMEs, que, muitas vezes, apresentam inconsistência nos dados financeiros e contábeis.

Altman et al. (2010) mostraram que pequenas e médias empresas são, em sua maioria, os maiores devedores de bancos e que boa parte das informações financeiras provindas destas empresas não estão disponíveis facilmente ou podem ser inconsistentes. Os autores concluem que as pequenas e médias empresas devem ser tratadas de maneira distinta, e que aprimorar os modelos de previsão com dados não financeiros faz com que a capacidade de previsão dos modelos de *credit scoring* aumente substancialmente. Ortiz-Molina e Penas (2008) estudaram as restrições de crédito para o segmento de MPMEs nos Estados Unidos e no Reino Unido. Esses autores associaram essas restrições à problemas de informação assimétrica e maiores riscos.

Lugovskaya (2010), em seu estudo sobre a utilização de variáveis não financeiras em pequenas e médias empresas, ilustra as desvantagens da utilização de demonstrações financeiras como a única fonte de informações para a previsão da inadimplência. O autor encontra, mediante a estimação de modelos, que o tamanho e a idade da empresa são variáveis significativas na previsão.

Conforme a Serasa Experian (2019), empresa que reúne a maior base de informações sobre crédito no Brasil, o número de empresas inadimplentes alcançou 5,7 milhões, em 2019, batendo o recorde histórico no país. As micro e pequenas empresas representaram 95% desse total. Houve um forte incremento do número de empresas inadimplentes durante a crise econômica brasileira que ocorreu entre 2014 e 2016. Em períodos de recessão, a

probabilidade de inadimplência de empresas e pessoas físicas aumenta consideravelmente devido à dificuldade de levantar recursos para honrar com os compromissos financeiros. Nesse contexto, a análise do risco de crédito é essencial para a tomada de decisões corporativas e para a prosperidade dos negócios (Sharpe et al., 1998).

## 2.2 Estudos prévios

Há uma ampla literatura sobre a utilização de métodos estatísticos para a previsão da inadimplência. A regressão logística é o método tradicionalmente utilizado nessa avaliação. Contudo, a literatura tem evoluído no sentido da adoção de métodos mais sofisticados como os de *machine learning*.

Araújo e Carmona (2007), em sua análise sobre a aplicação de *credit scoring* em uma instituição de microcrédito, mostraram que, mesmo em situações em que a modalidade de crédito é diferenciada, é possível a utilização da regressão logística para previsão da inadimplência. O estudo desses autores reforça a importância da análise estatística para a sustentabilidade financeira das instituições de microcrédito.

A análise de variáveis não econômicas para previsão de inadimplência de MPMEs vem sendo tema de diversos estudos sobre sobrevivência e inadimplência de empresas. Camargos et al. (2010) voltaram sua análise para a inadimplência deste segmento em uma instituição financeira pública estadual. Esses autores utilizaram a regressão logística para identificar quais os fatores condicionantes ao não pagamento de financiamentos.

Gonçalves et al. (2013) utilizaram um modelo de regressão logística, com 28 variáveis e três amostras, para a avaliação de concessão de crédito, confirmando a importância da utilização de métodos de previsão para avaliação de risco de crédito acordo com o apetite de risco da empresa. Eles destacaram que os modelos podem ser aprimorados com outras variáveis cadastrais dos clientes.

Mais recentemente, a literatura sobre *credit scoring* tem mudado seu foco para modelos com capacidades preditivas cada vez maiores. Silvério (2015) ressaltou a necessidade de padronização e agilidade nas análises de crédito. Segundo o autor, em análises com grande quantidade de informações com assertividade, algoritmos de *machine learning* têm apresentado resultados superiores aos da regressão logística. Os resultados desse autor sugerem o melhor desempenho do *random forest* sobre a regressão logística. Silvério (2015)

destaca também a importância da estabilidade do *random forest* para *credit scoring*, visto que algoritmos de *machine learning* conseguem lidar com a heterogeneidade de dados e estruturas, portanto, capturam bem ruídos ou não linearidades/irregularidades.

Brito Filho e Artes (2018) avaliaram os modelos de *machine learning Bayesian Additive Regression Trees* (*BART*) e *random forest* na previsão de *bons* ou *maus* pagadores, adotando, ainda, o modelo de regressão logística, por ser, segundo Thomas (2009), o modelo mais utilizado no mercado, como *benchmarking* para desempenho. Eles concluíram que, apesar de os modelos de *machine learning* não obterem ganhos significativos quando comparados ao de regressão logística, a pequena diferença de desempenho pode gerar ganhos financeiros significativos, quando indicar a não concessão de empréstimo para um *mau* pagador, ou pelo aumento de receita, quando indicar a concessão de empréstimo a um *bom* pagador.

Xiaojiao (2017) destaca que informações de redes sociais são importantes na previsão de inadimplência. O autor realizou uma comparação dos modelos *random forest* e *XGBoost* para entender qual apresenta melhor capacidade de prever a inadimplência em empréstimos bancários. Mediante a estatística de Kolmogorov-Smirnov (KS) ele encontrou resultados que sugerem que o método *XGBoost* apresenta melhor desempenho na previsão de inadimplência.

Segundo Forti (2018), a utilização de técnicas de *machine learning* por empresas tem aumentado nos últimos anos. Em seu estudo comparativo dos algoritmos *random forest*, *Support Vector Machine* e *gradient boosting*, que buscava identificar qual é capaz de prever melhor quais clientes estão mais propensos a pagar suas dívidas, a autora observou o melhor poder de predição das técnicas de *machine learning* frente à regressão logística. Ancieto (2016), em sua revisão sobre *machine learning*, argumenta que os métodos *Support Vector Machine*, *decision tree*, *bagging*, *AdaBoost* e *random forest* são mais eficientes que o método tradicional.

Becker (2018), em sua comparação de desempenho de diferentes modelos para concessão de crédito, sugere o desempenho superior de algoritmos de *machine learning* frente à regressão logística. Porém considera questionáveis os ganhos na utilização de algoritmos de *machine learning*, ao analisar o desempenho e a facilidade de implementação dos modelos logísticos.

Hamori e Kume (2018) compararam a capacidade preditiva dos métodos *XGBoost* e Redes Neurais. Ao prever o crescimento da economia de 134 países, sem a exclusão de *missings* — um importante fator que confere vantagem quando adotado o *XGBoost*, ao invés do *random forest* e da regressão logística. Esses autores mostraram que o *XGBoost* obtém melhores resultados que os modelos testados de redes neurais.

Marra (2019) reforça a importância de prever dificuldades financeiras na concessão de crédito e como a aplicação do *machine learning* tem corroborado para melhoria do processo decisório e avaliação de risco. Em seu estudo, no qual foram analisados índices financeiros de empresas latino-americanas, o *XGBoost* supera de forma incremental o modelo *random forest* e de forma expressiva a regressão logística. Isso leva a concluir que a evolução da capacidade preditiva do *machine learning* vem trazendo ganhos substanciais para a análise do risco de crédito.

## 3 MÉTODOS

### 3.1 Regressão logística

A regressão logística começou a ser utilizada com mais frequência para analisar crédito a partir da década de 1980 (Thomas, 2009). Desde então, se tornou um dos modelos mais utilizados na hora de se avaliar crédito no mercado. Por 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.

Assim, na regressão logística, a variável dependente, uma vez que possui caráter não-métrico, é representada mediante variáveis *dummies* (dicotômicas ou binárias), que, normalmente, assumem valor 0 para indicar a ausência de um atributo e 1 para indicar a presença de um atributo (Gujarati, 2000).

No contexto deste estudo, seja $\hat{p}_i = P(Y_i = 1)$, a probabilidade de um cliente ser um *mau* pagador, o modelo de regressão logística é dado pela Equação 1:

$$p_i = \frac{e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi}}}{1 + e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi}}}, \quad (1)$$

podendo ser linearizado por meio da transformação do logito, dada pela Equação 2:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}. \quad (2)$$

A variável dependente é a situação de adimplência do cliente, assumindo valor 1, quando ele é inadimplente, e valor 0, quando ele é adimplente. As variáveis independentes representam os fatores que podem explicar a inadimplência, podendo ser dados pessoais, financeiros, econômicos, entre outros. Os coeficientes β representam medidas de variações nas probabilidades conforme as variáveis independentes. Para estimar esses coeficientes, utiliza-se o método de máxima verossimilhança (Hosmer & Lemeshow, 2000).

Corrar et al. (2007) destacam a regressão logística pela possibilidade de contornar restrições existentes em outros modelos multivariados. Porém, ela é sensível à multicolinearidade entre as variáveis (Hair et al., 2005), a qual pode ser avaliada pelo o fator de inflação da variância (VIF).

### 3.2 *Machine learning*

O *machine learning* foi definido por Arthur Samuel, no final dos anos 50, como um campo de estudo que garante aos computadores a capacidade de aprender de forma autônoma após serem programados para essa tarefa. Ele pode ser entendido como uma subcategoria da inteligência artificial, um sistema que é capaz de analisar uma grande quantidade de dados através da estatística, identificando os padrões existentes para posteriormente fazer predições ou determinar relações (Pimentel & Omar, 2006).

Com a melhora na capacidade de processamento de computadores e a enorme quantidade de dados existentes atualmente, decisões que anteriormente eram tomadas qualitativamente se tornam cada vez menos importantes no mundo corporativo. Hoje, a presença de mecanismos confiáveis, que são capazes de criar relações e aprender continuamente, é uma realidade na maioria das empresas que concedem crédito. A contínua melhoria da capacidade preditiva dos modelos pela própria utilização dos mesmos é um pressuposto fundamental do *machine learning* (Silvério, 2015).

Existem dois tipos principais de aprendizagem: i) supervisionada, na qual um conjunto de dados é utilizado e se tem um resultado esperado; e ii) não supervisionada, na qual não é possível prever os resultados do cruzamento das informações. Segundo Reed e Marks (1999), a vantagem da aprendizagem supervisionada é que seu modelo é bem definido. Por outro

lado, grande vantagem da aprendizagem não supervisionada é poder ser utilizada de maneira mais abrangente, em virtude de dados não legendados se encontrarem frequentemente em maior disponibilidade que os dados classificados.

### 3.2.1 Métodos *ensemble*

Métodos *ensemble* são algoritmos que utilizam modelos simples, de baixo poder preditivo, combinados para resultar em um modelo mais potente, com maior acurácia. Estes se dividem em duas abordagens metodológicas: o *bagging* e o *boosting.* As principais diferenças e características dos modelos *ensemble* são apresentadas na Figura 1:

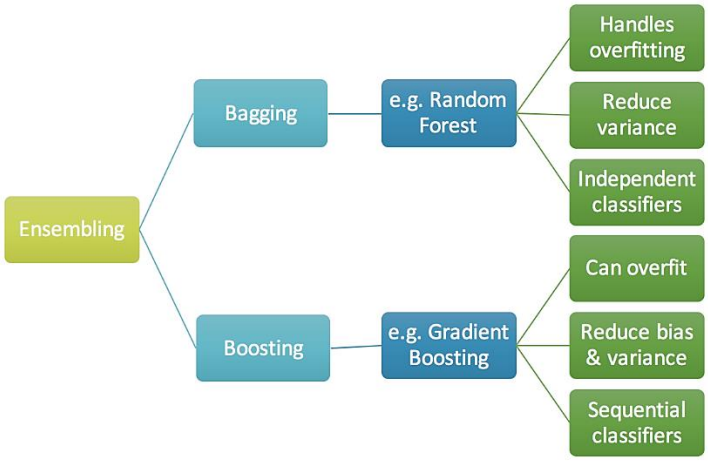



**Figura 1.** Diferentes Métodos Ensemble
Fonte: Grover (2017).

Nos algoritmos *bagging* os classificadores são treinados separadamente e agregados posteriormente por algum método de agregação, como, por exemplo, pela média. Por outro lado, nos algoritmos *boosting*, que também são treinados separadamente, a agregação é feita por uma ponderação do desempenho de cada modelo. Dentro de algoritmos *bagging* o mais conhecido é o *random forest* e, no caso do *boosting*, o *gradient boosting*.

#### 3.2.1.1 *Random forest*

O *random forest* é um algoritmo *ensemble* de *machine learning* que se baseia em árvores de decisão (*decision trees).* Esse algoritmo, proposto por Breiman (2001), é um aprimoramento do método *bagging*, utilizando uma seleção aleatória de apenas algumas das variáveis independentes na construção das árvores de decisão. Esse algoritmo consegue

reduzir a correlação entre os erros gerados em cada árvore, mantendo a força individual de tais árvores.

A principal vantagem do *random forest* é reduzir a correlação entre os erros gerados nas árvores de decisão do método *bagging*, sem aumentar muito a variância através da seleção aleatória das variáveis de entrada. Nesse método, considerando cada árvore de decisão gerada como independentemente distribuída, a média esperada de B árvores de decisão é a mesma esperada para qualquer uma delas (Hastie et al., 2001).

Segundo Breiman (2001), *random forest* é um classificador que consiste em uma coleção de árvores, estruturadas com k vetores aleatórios independentes e identicamente distribuído. O Apêndice do presente estudo traz os procedimentos de um algoritmo *random forest*, assim como uma ilustração do particionamento de uma árvore de decisão de *credit scoring*.

#### 3.2.1.2 *Extreme Gradient Boosting*

O algoritmo *assemble* chamado *gradient boosting* consiste em uma generalização do método de *AdaBoost* proposto por Freund e Schapire (1996). Usualmente, o método *gradient boosting* utiliza o modelo de *decision trees* como base. Porém, outros métodos de regressão podem ser utilizados em sua aplicação.

A ideia principal de todo algoritmo em *boosting* é minimizar a função perda em cada árvore de decisão. A função objetivo é continuamente minimizada a partir da utilização de múltiplas combinações de árvores aleatoriamente estimadas, até se atingir o mínimo global, que minimiza o erro de estimação. O algoritmo funciona como a combinação de diversas regras de aprendizagens fracas que ao serem combinadas são convertidas em um modelo forte.

O *Extreme gradient boosting* foi desenvolvido por Chen e Guestrin (2016) e se difere do *gradient boosting* no que diz respeito a otimizar a função objetivo de maneira mais robusta, auxiliando na regularização para controlar o ajuste (*overfitting*), de forma que o modelo tenha alta precisão quando testado em seu conjunto de dados. Porém o método perde eficácia na previsão de novos resultados.

O algoritmo funciona da seguinte maneira: primeiramente cria-se um modelo inicial ($F_0$) que prediz a variável alvo; um novo modelo ($h_1$) é criado para complementar o valor

residual do primeiro modelo ($Y - F_0(x)$), sendo Y os valores reais e $F_0(x)$ os valores preditos; então, os modelos $F_0$ e $h_1$ são combinados gerando o modelo $F_1$, a *boosted version* de $F_0$. O modelo F1 é dado pela Equação 3:

$$F_1(x) = F_0(x) + h_1(x) \tag{3}$$

A média de erro quadrático de $F_1$ é menor do que de $F_0$, o que indica que o modelo erra menos. Para melhorar o modelo $F_1$ basta gerar um modelo $h_2$ que complementa os valores residuais de $F_1(Y - F_1(x))$ e combiná-los gerando um modelo $F_2$. Este processo se repete até que o valor residual seja reduzido o máximo possível.

Desde sua criação, este método tem sido considerado um dos mais robustos de *machine learning*. Até então, ele foi o mais vitorioso no Kaggle, uma comunidade *online* de competições em que cientistas de dados resolvem desafios por meio de *machine learning*.

**3.3 Métricas de validação**

Para avaliar o desempenho dos métodos analisados neste estudo, foram utilizados o teste de Kolmogorov-Smirnov (KS) e a acurácia (ACC) dos modelos. Apesar de não serem comuns em estudos que se utilizam de métodos estatísticos convencionais, tais métricas são amplamente utilizadas quando se trata da avaliação de modelos de *machine learning*.

A ACC é uma medida da precisão do modelo em comparação aos dados gerais. Ela é calculada como a razão entre as unidades corretamente classificadas, verdadeiro positivo (TP) e falso positivo (FP), e o número total de previsões feitas pelos classificadores, considerando os falsos negativos (FN) e os verdadeiros negativos (TN). ACC pode ser calculada conforme a Equação 4:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}. \tag{4}$$

O KS é usado para medir a adequação do modelo. Ele indica a distância entre a função de distribuição empírica de uma amostra e a função de cumulativa de uma distribuição de referência. Esse teste também pode ser usado para comparar distribuições de dois conjuntos de amostra, em vez de um conjunto de amostra e uma distribuição de referência. Considerando TP e FP, no contexto deste estudo, o KS pode ser representado de acordo com a Equação 5:

$$\mathrm{KS} = \max_t(|\mathrm{TP}_t - \mathrm{FP}_t|). \tag{5}$$

O KS é um dos métodos mais difundidos para comparação de amostra dos modelos de *machine learning* (Zhang et al., 2018).

## 4 DADOS E VARIÁVEIS

O objetivo deste estudo é comparar diferentes métodos estatísticos na predição da inadimplência. Considerou-se como inadimplentes clientes que ficaram com títulos em aberto em um prazo maior que 90 dias. Os clientes adimplentes foram classificados como *Bons* e os clientes inadimples foram classificados como *Ruins*. A base de clientes utilizada foi cedida por uma empresa do setor serviços de locação de automóveis. Ela agrega informações próprias, bem como aquelas da Serasa Experian. Os dados são em corte transversal e as informações disponibilizadas são do período de janeiro de 2018 a abril 2019. A amostra inicial de 17.800 observações possui clientes de todo o território nacional. Essa amostra inicial foi balanceada aleatoriamente, no que diz respeito à quantidade de clientes classificados como *Bons* e *Ruins.* A amostra final continha 3.844 clientes, dos quais 1.492 eram classificados como *Bons* e 1.492 eram classificados como *Ruins*.

Foram definidas como *variáveis internas* todas aquelas relacionadas aos dados financeiros e contratuais dos clientes levantados pela empresa que concedeu os dados, além daquelas de cadastro do cliente. Foram definidas como *variáveis externas* aquelas de agências de crédito. As variáveis *collection* e *credit score* consistem em modelos de classificação dos clientes utilizados ao nível de mercado. *Collection score* indica a probabilidade de o cliente pagar seus débitos. *Credit score* indica a probabilidade de o cliente ficar inadimplente. Mesmo existindo tais métricas para avaliação dos clientes ao nível de mercado, os estudos que empregam métodos estatísticos auxiliam no aprofundamento da análise da concessão de crédito aos clientes, considerando variáveis específicas para o nicho de atuação da empresa.

**Tabela 1**
Variáveis internas

| Variável | Descrição | Tipo |
|---|---|---|
| Limite de Crédito | Variável categórica do valor que determina a quantidade de carros que podem ser alugados pelo cliente, solicitado em intervalos de R$ 5 mil em R$ 5 mil até o limite final de R$ 25 mil. | Categórica, com 5 categorias |
| Regional de Vendas | Variável categórica das regionais do território brasileiro. | Categórica, com 11 categorias |
| Quantidade de Contratos | Quantidade de contratos do cliente com a empresa: identifica a quantidade de aluguéis do cliente. | Contínua |

| Variável | Descrição | Tipo |
|---|---|---|
| Quantidade de Contratos Multa | Quantidade de contratos com multa: identifica a quantidade de multas do cliente. | Contínua |
| Quantidade de Contratos Avaria | Quantidade de contratos com avaria: identifica a quantidade de vezes em que o cliente danificou veículos. | Contínua |
| Valor Total das Faturas Pagas à Vista | Valor total das vendas à vista: identifica o valor que o cliente pagou à vista. | Contínua |
| Valor Total das Faturas Pagas a Prazo | Valor total das vendas a prazo: identifica quanto o cliente pagou a prazo. | Contínua |
| Valor Total das Faturas a Prazo em Aberto | Valor total das faturas em aberto: valor da dívida do cliente. | Contínua |
| Valor Total Cancelado | Valor total das faturas canceladas. | Contínua |

Fonte: Elaboração própria.

**Tabela 2**
Variáveis externas

| Variável | Descrição | Tipo |
|---|---|---|
| CNAE | Classificação Nacional de Atividades Econômicas, agregado em: Comércio, Engenharia, Indústria e Serviços. | Categórica, com 4 categorias |
| Capital Social | Valor do capital social da empresa. | Contínua |
| Percentual de Pagamentos Pontuais | Percentual das dívidas da empresa pagas pontualmente no mercado. | Contínua |
| Idade da Empresa | Idade da empresa em anos. | Contínua |
| Inadimplência dos Sócios | Indicação de inadimplência dos sócios (pessoa física) junto ao mercado. | Binária |
| Número de Filiais | Número de filiais do CNPJ em que foi solicitado a análise. | Contínua |
| Score Cobrança | Nota do *collection score* de mercado. | Contínua |
| Score de Crédito | Nota do *credit score* de mercado. | Contínua |

Fonte: Elaboração própria.

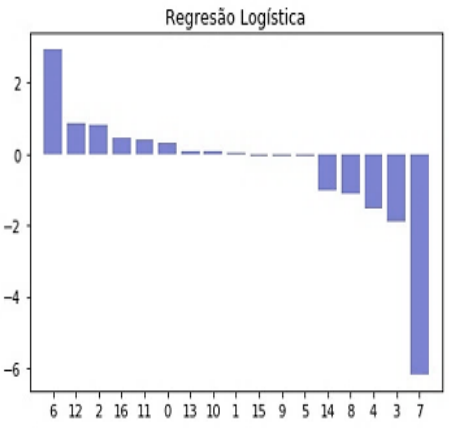
As tabelas 1 e 2 apresentam, respectivamente, as variáveis internas e externas utilizadas neste estudo. É importante destacar que nelas estão incluídas tanto variáveis financeiras quanto variáveis não financeiras. Altiman et al. (2010) destacam que as variáveis não financeiras podem incrementar substancialmente a capacidade preditiva dos modelos de *credit scoring*.

## 5 RESULTADOS

Os resultados deste estudo são baseados em três modelos preditivos estimados em Python. Para a comparação dos resultados obtidos nos modelos de *credit scoring*, foram utilizadas as métricas: ACC e KS. O peso das variáveis dentro de cada modelo preditivo foi ordenado, a fim de identificar quais delas impactam mais na classificação dos clientes como inadimplente ou adimplente. Além disso, foi gerada a matriz de confusão e a Curva ROC a fim de identificar a capacidade de previsão de verdadeiros/falsos positivos e negativos.

### 5.1 Resultados da regressão logística

No modelo tradicional de regressão logística, não é possível a utilização direta de variáveis categóricas. Para isso, é necessária a criação de diversas variáveis *dummies* que dizem respeito a Limite de Crédito, Regional de Vendas e CNAE.




| Ranking de Carcterísticas | | |
|---|---|---|
| 1 | Valor Total das Faturas a Prazo em Aberto | -6,1834 |
| 2 | Valor Total das Faturas Pagas a Prazo | 2,9215 |
| 3 | Quantidade de Contratos Multa | -1,8888 |
| 4 | Quantidade de Contratos Avaria | -1,5259 |
| 5 | Valor Total Cancelado | -1,1279 |
| 6 | Número de Filiais | -1,0200 |
| 7 | Capital Social | 0,8671 |
| 8 | Quantidade de Contratos | 0,8171 |
| 9 | Score de Crédito | 0,4285 |
| 10 | Idade da Empresa | 0,3895 |
| 11 | Limite de Crédito | 0,3225 |
| 12 | Inadimplência dos Sócios | 0,0833 |
| 13 | Percentual de Pagamentos Pontuais | 0,0766 |
| 14 | CNAE | -0,0415 |
| 15 | Score Cobrança | -0,0389 |
| 16 | Regional de Vendas | 0,0350 |
| 17 | Valor Total das Faturas Pagas à Vista | -0,0084 |

**Figura 2.** Peso das Variáveis na Regressão Logística
Fonte: Elaboração própria

Os resultados do modelo estimado mostram que variáveis como *valor das faturas a prazo pago* e *o valor das faturas a prazo em aberto* são de grande importância no momento de classificar os clientes como *Bons* ou *Ruins*. A Figura 3 mostra a acurácia alcançada pelo modelo dividido entre as bases de treinamento.

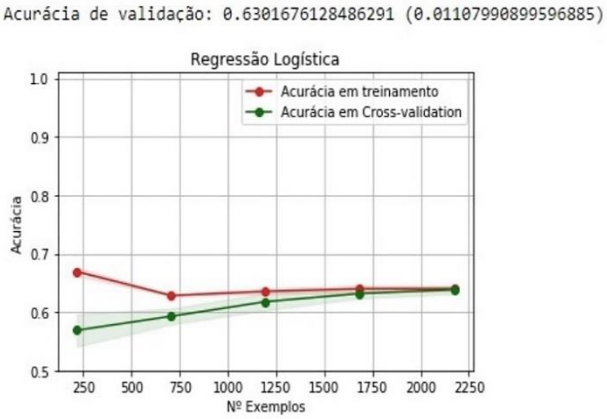



**Figura 3.** Acurácia da Regressão Logística nas Bases de Treinamento e Validação
**Nota.** Acurácia de validação: 0,6302 (0,0111). Fonte: Elaboração própria.

Neste estudo, foram utilizadas como base de treinamento uma amostra com 2.718 observações e, para a validação, uma amostra com 1.166 observações, mostrando que um maior número de observações aumenta a acurácia do modelo. Foi utilizado o método de validação cruzada *k-fold cross-validation* (Kohavi, 1995) para avaliar a capacidade de generalização do modelo. Neste método, os dados foram aleatoriamente particionados em 5 grupos, *datasets* (k = 5), e, posteriormente, foi estimada a média da acurácia de cada um destes *datasets* para se ter a acurácia final do modelo. A Figura 4 mostra a matriz de confusão do modelo, que apresenta o cruzamento das classificações reais *versus* as preditas, sendo uma maneira prática para a visualização dos positivos reais frente aos falsos positivos, e o mesmo para os negativos.

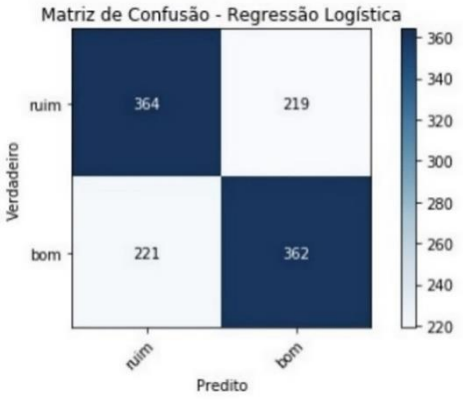


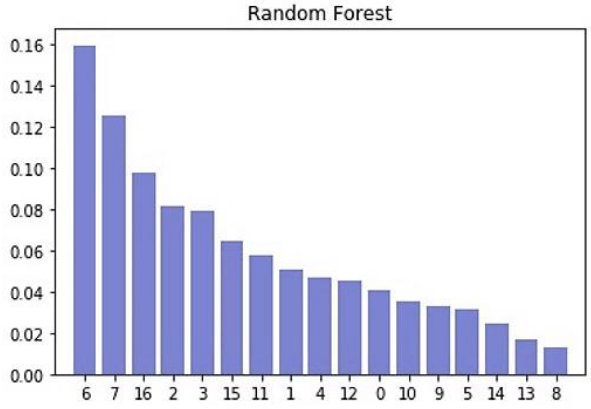

**Figura 4.** Matriz de Confusão da Regressão Logística
Fonte: Elaboração própria.

O modelo de regressão logística, na base de validação, foi capaz de prever corretamente 364 clientes classificados como *Ruins* corretamente e 221 de maneira incorreta, assim como

previu 362 clientes classificados como *Bons* de maneira correta e 219 desses clientes de maneira incorreta. Logo, o modelo foi capaz de prever 62% das observações de maneira correta. Silvério (2015), em seu estudo sobre a aplicação de diferentes métodos para análise de crédito, traz um resultado semelhante em relação a capacidade preditiva do modelo de regressão logística ao prever corretamente 64% de suas observações.

### 5.2 Resultados do *random forest*

O *random forest* estimado neste estudo, conforme seu algoritmo padrão, emprega dez árvores de decisão. Os resultados desse modelo são exibidos na Figura 5. Diferentemente do modelo de regressão logística, os modelos de árvore de decisão utilizam equação probabilística para criar seus nós e, dessa forma, indicar a relevância das variáveis, o que é denotado conforme valores que estão sempre entre 0 e 1.




| | Ranking de Carcterísticas | |
|---|---|---|
| 1 | Valor Total das Faturas Pagas a Prazo | 0,1592 |
| 2 | Valor Total das Faturas a Prazo em Aberto | 0,1249 |
| 3 | Score de Crédito | 0,0977 |
| 4 | Quantidade de Contratos | 0,0810 |
| 5 | Quantidade de Contratos Multa | 0,0792 |
| 6 | Score Cobrança | 0,0641 |
| 7 | Idade da Empresa | 0,0573 |
| 8 | Regional de Vendas | 0,0509 |
| 9 | Quantidade de Contratos Avaria | 0,0467 |
| 10 | Capital Social | 0,0449 |
| 11 | Limite de Crédito | 0,0409 |
| 12 | Percentual de Pagamentos Pontuais | 0,0349 |
| 13 | CNAE | 0,0328 |
| 14 | Valor Total das Faturas Pagas à Vista | 0,0315 |
| 15 | Número de Filiais | 0,0246 |
| 16 | Inadimplência dos Sócios | 0,0164 |
| 17 | Valor Total Cancelado | 0,0129 |

**Figura 5.** Peso das Variáveis no Random Forest
Fonte: Elaboração própria.

Assim como no modelo de regressão logística, as variáveis mais importantes são *o valor das faturas a prazo pago* e *o valor das faturas a prazo em aberto.* Todavia, como eles são de modelos distintos, no *random forest,* algumas variáveis têm maior importância para decisão do que no modelo de regressão logística. A Figura 6 mostra a análise da acurácia do modelo *random forest.*

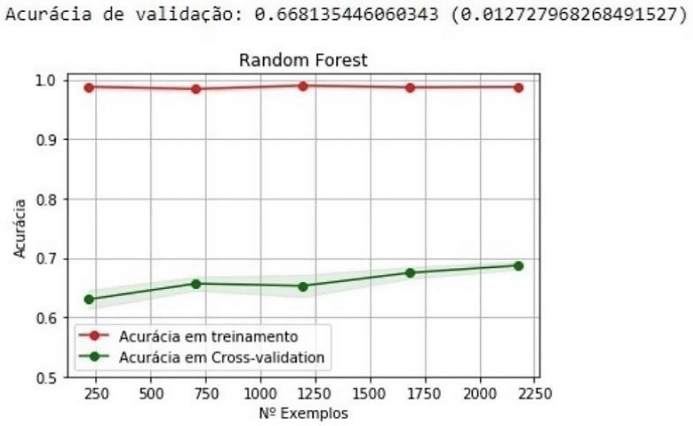



**Figura 6.** Acurácia do Random Forest nas Bases de Treinamento e Validação
**Nota.** Acurácia de validação: 0,6681 (0,0127). Fonte: Elaboração própria.

A Figura 6 mostra que na base de treinamento o modelo se manteve com uma acurácia bem mais alta e constante. Isso provavelmente ocorre quando o algoritmo recebe exemplos de mesma classe em sequência. Por outro lado, na base de validação há exemplos que não foram considerados pelo modelo. Portanto, é esperado que a acurácia dele seja menor e cresça conforme aumenta a quantidade de observações.

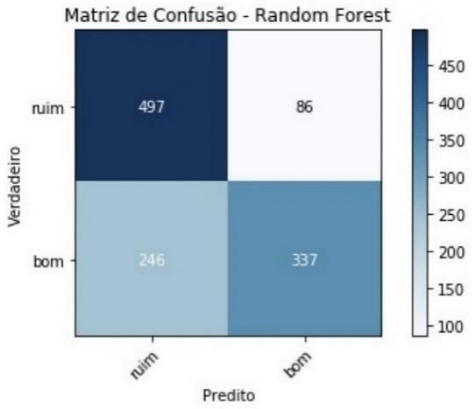


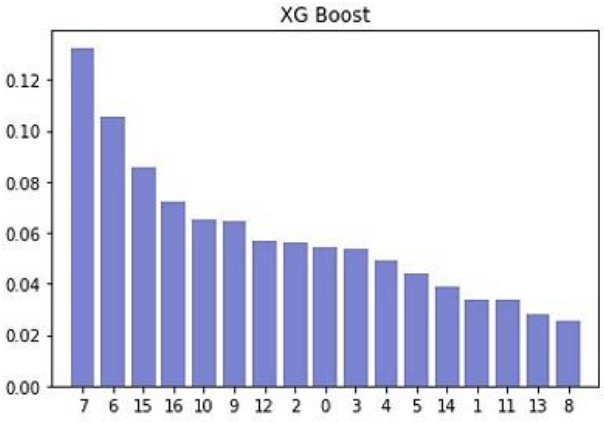

**Figura 7.** *Matriz de Confusão do* Random Forest
Fonte: Elaboração própria.

A matriz de confusão do *random forest*, mostrada na Figura 7, feita conforme a base de validação, foi capaz de prever corretamente 497 clientes classificados como *Ruins* corretamente e 246 de maneira incorreta, assim como previu 337 clientes classificados como *Bons* de maneira correta e 86 incorretamente. Logo, método *random forest* foi capaz de prever corretamente 72% da amostra estudada para validação. Isso mostra que o modelo tem uma boa capacidade para prever os clientes da classe *Ruins* de maneira correta. O resultado obtido

é um pouco superior ao resultado de 69% na base de validação, obtido por Becker (2018) ao utilizar variáveis não financeiras para previsão de crédito.

### 5.3 Resultados do *XGBoost*

Assim como o *random forest,* o *XGBoost* é baseado em um conjunto de árvores. Porém, o algoritmo padrão cria 100 árvores para a decisão. Os resultados obtidos com o modelo *XGBoost,* são apresentados na Figura 8.




| | Ranking de Carcterísticas | |
|---|---|---|
| 1 | Valor Total das Faturas a Prazo em Aberto | 0,1326 |
| 2 | Valor Total das Faturas Pagas a Prazo | 0,1057 |
| 3 | Score Cobrança | 0,0858 |
| 4 | Score de Crédito | 0,0722 |
| 5 | Percentual de Pagamentos Pontuais | 0,0650 |
| 6 | CNAE | 0,0642 |
| 7 | Capital Social | 0,0572 |
| 8 | Quantidade de Contratos | 0,0565 |
| 9 | Limite de Crédito | 0,0545 |
| 10 | Quantidade de Contratos Multa | 0,0535 |
| 11 | Quantidade de Contratos Avaria | 0,0493 |
| 12 | Valor Total das Faturas Pagas à Vista | 0,0439 |
| 13 | Número de Filiais | 0,0388 |
| 14 | Regional de Vendas | 0,0340 |
| 15 | Idade da Empresa | 0,0336 |
| 16 | Inadimplência dos Sócios | 0,0277 |
| 17 | Valor Total Cancelado | 0,0257 |

**Figura 8.** Peso das Variáveis no XGBoost
Fonte: Elaboração própria.

O modelo *XGBoost* também é estimado mediante árvores de decisão, portanto, ele é probabilístico. Seguindo o padrão dos modelos anteriores, as variáveis que tiveram maior relevância foram os *valores das faturas a prazo pago* e *valor das faturas a prazo em aberto*. É importante destacar o alto peso dos dois *scores* de mercado (c*ollection score* e *credit score*) na composição da decisão. Essas variáveis não financeiras também parecem importantes para a previsão de inadimplência. Isso está em linha com a perspectiva de Altman et al. (2010), que afirmam que variáveis não financeiras são importantes para o *credit scoring*.

A acurácia do modelo é mostrada na Figura 9. Ela indica que a cada novo *dataset* o modelo se torna mais confiável, reafirmando a necessidade de se ter mais observações para que o modelo tenha uma melhor acurácia.

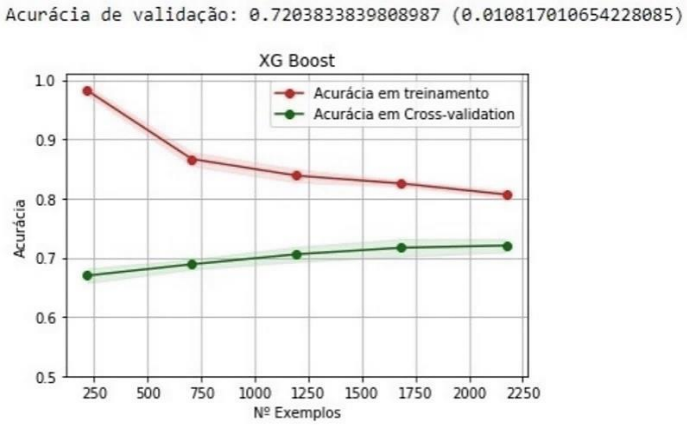



**Figura 9.** Acurácia do XGBoost nas Bases de Treinamento e Validação
**Nota.** Acurácia de validação: 0,7203 (0,0108). Fonte: Elaboração própria.

A matriz de confusão do *XGBoost*, mostrada na Figura 10, indica que o modelo foi capaz de prever corretamente 480 clientes classificados como *Ruins* corretamente e 164 deles de maneira incorreta, assim como previu 419 clientes classificados como *Bons* de maneira correta e 103 deles incorretamente.

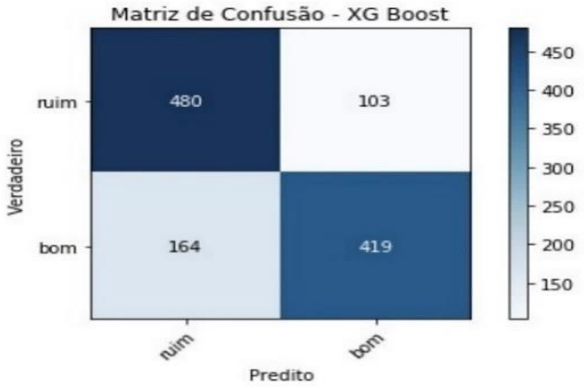



**Figura 10.** Matriz de Confusão do XGBoost
Fonte: Elaboração própria.

Portanto, o modelo foi capaz de prever 77% das observações de maneira correta. Esse resultado está em conformidade com os estudos recentes feitos por Xiaojiao (2017) e Hamori e Kume (2018), que também identificaram uma maior capacidade preditiva do método *XGBoost*.

### 5.4 Comparação entre os modelos

A Curva ROC fornece uma medida de precisão total independente, ao ilustrar o desempenho de um sistema classificador binário à medida que seu limiar de discriminação varia, revelando de maneira centralizada que os modelos de *machine learning* estudados tiveram maior capacidade preditiva frente à regressão logística. Os valores da área abaixo da diagonal (50%) não têm validade, por serem considerados como aleatórios. Quanto mais próximo de 100% melhor é a capacidade do modelo fazer previsões corretas. A Figura 11 apresenta a Curva ROC dos modelos testados e sugere que o modelo estimado pelo método *XGBoost* teve mais verdadeiros positivos que os outros modelos.

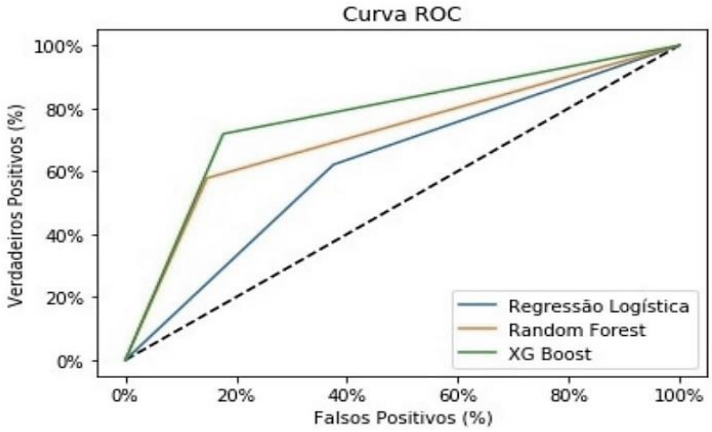



**Figura 11.** Curva ROC
Fonte: Elaboração própria.

Finalmente, para medir o desempenho dos modelos estudados foi analisado o KS e a ACC. A Tabela 3 exibe uma grande diferença dos resultados dos modelos de *machine learning* frente ao modelo de regressão logística, sobretudo, quando observado o KS. Pode-se concluir que o *XGBoost* tem maior capacidade preditiva entre os métodos analisados, tendo em vista os maiores valores tanto do KS quanto da ACC. Marra (2019) utilizou os mesmos métodos na previsão de dificuldades financeiras em empresas latino-americanas. Esse autor chegou a resultados semelhantes quanto à ordem de classificação da capacidade preditiva dos modelos estimados pelos distintos métodos estatísticos.

**Tabela 3**
Desempenho dos Modelos

| Métodos | KS | ACC |
|---|---|---|
| Regressão logística | 30,36% | 63,02% |
| *Random forest* | 47,00% | 66,81% |
| *XGBoost* | 57,12% | 72,04% |

Fonte: Elaboração própria.

Observa-se que quanto mais sofisticado o método estatístico adotado, maior foi a capacidade preditiva do modelo. Os resultados deste estudo sugerem que o método *XGBoost* é capaz de realizar melhores previsões do risco de inadimplência. Os algoritmos de árvores de decisão podem apresentar boas provisões, inclusive, em diferentes contextos, uma vez que se adaptam de maneira automática às mudanças ocorridas ao longo do tempo.

À título de conclusão, este estudo mostrou que o *XGBoost* apresentou maior capacidade preditiva quando comparado com os modelos estimados por outros métodos (*random forest* e regressão logística), levando-se em consideração os parâmetros utilizados.

## 6 CONSIDERAÇÕES FINAIS

Este artigo corrobora a necessidade da utilização de variáveis não financeiras para a previsão da inadimplência, particularmente, considerando uma amostra de MPMEs. Nele, utilizou-se de dados de clientes que tiveram solicitação de crédito aprovada. Logo, como uma limitação, não foi possível analisar os clientes que foram reprovados.

O resultado do modelo estimado pelo método *XGBoost*, quando comparado àqueles dos outros modelos, corrobora sua maior capacidade de previsão de inadimplência. Assim, este estudo confirma a superioridade dos métodos de *machine learning* frente à regressão logística. A evolução em termos computacionais e de sofisticação dos métodos faz com que seja importante um constante estudo dos melhores métodos de previsão, considerando aqueles de *machine learning*, que são cada vez mais empregados. Este estudo corroborou a superioridade da capacidade preditiva desses métodos mais modernos frente a um método clássico.

A constatação deste estudo é importante, em linha com a perspectiva de Tsai et al. (2014), em virtude de pequenas melhorias na precisão da classificação de crédito poderem implicar grandes reduções no risco e nas perdas para empresas ou instituições financeiras, contribuindo, assim, para o aumento da concessão do crédito com qualidade, um dos pilares para garantir um crescimento econômico sustentável.

Estudos futuros podem considerar bases de dados que tenham informações sobre os clientes que tiveram o crédito reprovado. Ademais, sugere-se uma análise mais detalhada dos parâmetros utilizados para calibrar o algoritmo, bem como a melhor divisão dos dados, tanto para o treinamento quanto para testar a capacidade preditiva e a acurácia do modelo.

## REFERÊNCIAS


Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance, 23*(4), 589–609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

Altman, E. I., Sabato, G., & Wilson, N. (2010). The value of non-financial information in SME risk management. *The Journal of Credit Risk, 6*(2), 95–127. https://doi.org/10.21314/jcr.2010.110

Altman, E. I., & Saunders, A. (1997). Credit risk measurement: Developments over the last 20 years. *Journal of Banking & Finance, 21*(11-12), 1721–1742. https://doi.org/10.1016/s0378-4266(97)00036-8

Aniceto, M. C. (2016). *Estudo comparativo entre técnicas de aprendizado de máquina para estimação de risco de crédito* [Dissertação de Mestrado, Universidade de Brasília]. https://doi.org/10.26512/2016.03.D.20522

Araújo, E. A., & Carmona, C. U. D. M. (2009). Desenvolvimento de modelos *credit scoring* com abordagem de regressão de logística para a gestão da inadimplência de uma instituição de microcrédito. *Contabilidade Vista & Revista, 18*(3), 107–131. Recuperado de https://revistas.face.ufmg.br/index.php/contabilidadevistaerevista/article/view/335

Becker, C. (2018). *Estudo comparativo entre abordagens de aprendizado de máquina em modelos de credit scoring* **[**Monografia de Graduação, Universidade Federal do Rio Grande do Sul]. Recuperado de http://hdl.handle.net/10183/201492

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Brito Filho, D. A., & Artes, R. (2018). Application of bayesian additive regression trees in the development of credit scoring models in Brazil. *Production, 28,* e20170110. https://doi.org/10.1590/0103-6513.20170110

Brito, G. A. S., & Assaf Neto, A. (2008). Modelo de classificação de risco de crédito de empresas. *Revista Contabilidade & Finanças, 19*(46), 18–29. https://doi.org/10.1590/S1519-70772008000100003

Camargos, M. A., Camargos, M. C. S., Silva, F. W., Santos, F. S., & Rodrigues, P. J. (2010). Fatores condicionantes de inadimplência em processos de concessão de crédito a micro e pequenas empresas do Estado de Minas Gerais. *Revista de Administração Contemporânea, 14*(2), 333–352. https://doi.org/10.1590/S1415-65552010000200009

Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. https://doi.org/10.1145/2939672.2939785

Corrar, L. J., Paulo, E., & Dias Filho, J. M. (2007). *Análise multivariada: Para cursos de administração, ciências contábeis e economia*. São Paulo: Atlas.

Durand, D. (1941). *Risk elements in consumer installment financing***.** New York: National Bureau of Economic Research.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics***,** *7*(2), 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Forti, M. (2018). *Técnicas de machine learning aplicadas na recuperação de crédito do mercado brasileiro.* [Dissertação de Mestrado, Fundação Getúlio Vargas]. Recuperado de http://hdl.handle.net/10438/24653

Freund, Y., & Schapire, R.E. (1996). Experiments with a new boosting algorithm. *ICML'96: Proceedings of the Thirteenth International Conference on Machine Learning*. Recuperado de https://cseweb.ucsd.edu/~yfreund/papers/boostingexperiments.pdf

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. New York: Springer.

Gonçalves, E. B., Gouvêa, M. A., & Mantovani, D. M. N. (2013). Análise de risco de crédito com o uso de regressão logística. *Revista Contemporânea De Contabilidade, 10*(20), 139–160. https://doi.org/10.5007/2175-8069.2013v10n20p139

Grover, P. (2017). *Gradient boosting from scratch*. Recuperado de https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d

Gujarati, D. N. (2000). *Econometria básica* (3a ed.). São Paulo: Makron Books.

Guégan, D., & Hassani, B. (2018). Regulatory learning: How to supervise machine learning models? An application to credit scoring. *The Journal of Finance and Data Science, 4*(3), 157–171. https://doi.org/10.1016/j.jfds.2018.04.001

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (2005). *Análise multivariada de dados* (5a ed.). Porto Alegre: Bookman.

Hamori, S., & Kume, T. (2018). Artificial intelligence and economic growth**.** *Advances in Decision Sciences, 22*(1), 256–278. https://doi.org/10.47654/v22y2018i1p256-278

Hand, D. J., & Adams, N. M. (2000). Defining attributes for scorecard construction in credit scoring. *Journal of Applied Statistics, 27*(5), 527–540. https://doi.org/10.1080/02664760050076371

Hosmer, D.W., & Lemeshow, S. (2000). *Applied logistic regression* (5a ed.). Danvers: John Wiley & Sons.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Appears in the International Joint Conference on Articial Intelligence (IJCAI)*. Recuperado de http://ai.stanford.edu/~ronnyk/accEst.pdf

Lewis, E. (1992). *Introduction to credit scoring.* San Rafael: Athena Press.

Lugovskaya, L. (2010). Predicting default of Russian SMEs on the basis of financial and non-financial variables. *Journal of Financial Services Marketing, 14*(4), 301–313. https://doi.org/10.1057/fsm.2009.28

Marra, V. N. (2019). *Previsão de dificuldades financeiras em empresas latino-americanas via aprendizagem de máquina***.** [Dissertação de Mestrado, Universidade Federal de Uberlândia]. http://dx.doi.org/10.14393/ufu.di.2019.947

Mitchell, T. M. (1997). *Machine learning*. New York: McGraw-Hill.

Munkhdalai, L., Munkhdalai, T., Namsrai, O.-E., Lee, J., & Ryu, K. (2019). An empirical comparison of machine-learning methods on bank client credit assessments. *Sustainability, 11*(3), 699. http://dx.doi.org/10.3390/su11030699

Ortiz-Molina, H., & Penas, M. F. (2007). Lending to small businesses: The role of loan maturity in addressing information problems. *Small Business Economics, 30*(4), 361–383. http://dx.doi.org/10.1007/s11187-007-9053-2

Pimentel, E., & Omar, N. (2006). Descobrindo conhecimentos em dados de avaliação da aprendizagem com técnicas de mineração de dados. *Anais do Workshop de Informática na Escola, 1*(1). Recuperado de https://www.br-ie.org/pub/index.php/wie/article/view/885

Qu, Y. (2008). Macroeconomic factors and probability of default. *European Journal of Economics, Finance and Administrative Sciences, 13*, 192–215.

Reed, R. D., & Marks, R. J. (1999). *Neuronal smithing: Supervised learning in feedward artificial neuronal network*. Cambridge: MIT Press.

Serasa Experian (2019). *Inadimplência de micro e pequenas empresas cresce 6,1% em maio, revela Serasa Experian.* Recuperado de https://www.serasaexperian.com.br/sala-de-imprensa/estudos-e-pesquisas/inadimplencia-de-micro-e-pequenas-empresas-cresce-61-em-maio-revela-serasa-experian/

Sharpe, W. F., Alexander, G. J., & Bailey, J. V. (1998). *Investments* (6a ed.). New Jersey: Prentice Hall.

Silverio, M. (2015). *Aplicação de algoritmos de aprendizado de máquina no desenvolvimento de modelos de escore de crédito***.** [Dissertação de Mestrado, Insper]. Recuperado de http://dspace.insper.edu.br/xmlui/handle/11224/1503

Sousa, A. F., & Chaia, A. J. (2000). Política de crédito: uma análise qualitativa dos processos das empresas. *Caderno de Pesquisas em Administração, 7*(3), 13-25.

Thomas, L. C. (2009). *Consumer credit models: Pricing, profit and portfolios*. New York: Oxford University Press.

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

Xiaojiao, Y. (2017). Machine learning application in online leading credit risk prediction, *ArXiv*. Recuperado de https://arxiv.org/abs/1707.04831

Zhang, L., Priestley, J., & Ni, X. (2018). Influence of the event rate on discrimination abilities of bankruptcy prediction models. *International Journal of Database Management Systems, 10*(1), 01–14. http://dx.doi.org/10.5121/ijdms.2018.10101

Zhong, H., Miao, C., Shen, Z., & Feng, Y. (2014). Comparing the learning effectiveness of BP, ELM, I-ELM, and SVM for corporate credit ratings. *Neurocomputing, 128*, 285–295. http://dx.doi.org/10.1016/j.neucom.2013.02.054

## Apêndice

Procedimentos do algoritmo *random forest:*

i) Para b = 1 até B:
   a) Amostra *bootstrap* Z de tamanho N a partir da base de treinamento;
   b) Selecione m variáveis aleatoriamente a partir de p variáveis;
   c) Selecione a melhor variável de acordo com seu poder de discriminação da variável de interesse;
   d) Divida o nó a partir de dois nós filhos;
ii) Reporte o *ensemble* de árvores.

Para fazer a previsão de um novo ponto x utilize a regressão: $\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$.

A Figura 12 ilustra o particionamento em uma árvore de decisão.

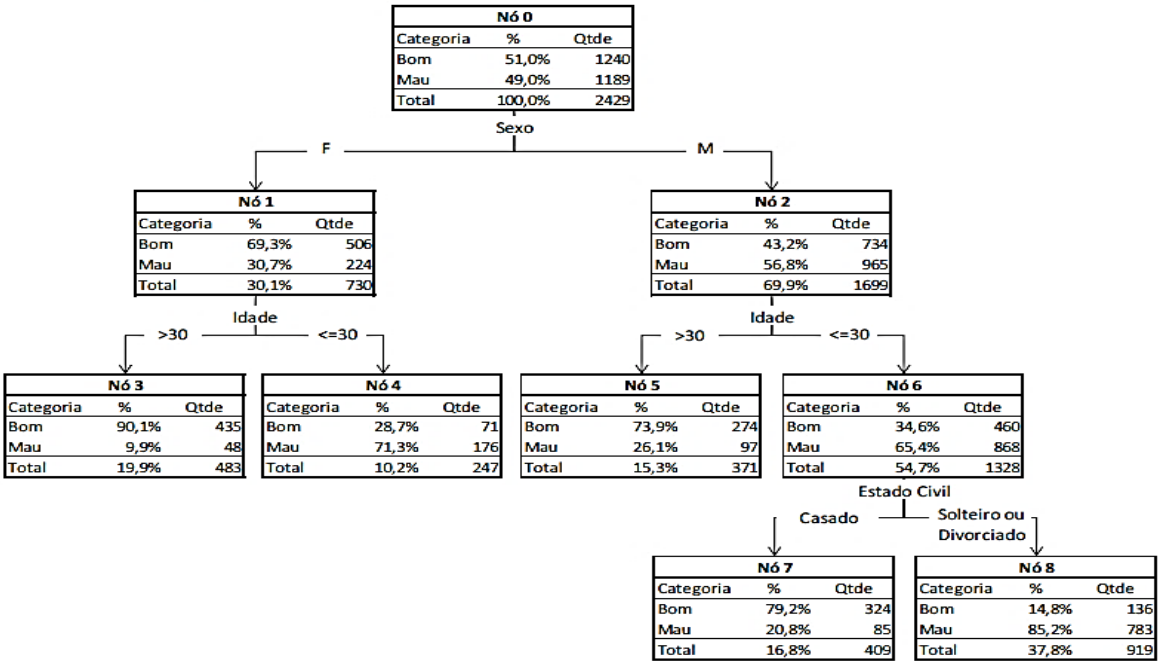



**Figura 12.** Ilustração de uma Árvore de Decisão
Fonte: Silvério (2015).