



Análise de Crédito Empregando Modelos de Aprendizado de Máquina

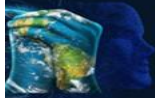
Luís Eduardo Moreira Las Casas
Universidade Federal de Minas Gerais
dudu_las-casas@hotmail.com
Renata Turola Takamatsu
Universidade Federal de Minas Gerais
rettakamatsu@gmail.com
Tâmara Leal Riveres
Universidade Federal de Minas Gerais
tamara.riveres@gmail.com

Resumo

O presente trabalho tem como objetivo avaliar o desempenho de algoritmos de aprendizado de máquina aplicados à análise de crédito, com foco na previsão de inadimplência de clientes. Diante do aumento expressivo da concessão de crédito e dos riscos associados à inadimplência, torna-se essencial que instituições financeiras adotem métodos mais robustos e automatizados para a tomada de decisão. Nesse contexto, investigou-se a aplicabilidade de três modelos preditivos Regressão Logística, Decision Tree e XGBoost utilizando uma base de dados pública da plataforma LendingClub, composta por mais de dois milhões de registros de solicitações de empréstimos realizadas entre 2007 e 2018. Após a seleção e o pré-processamento das variáveis, os modelos foram avaliados por meio das métricas Acurácia, Recall, F1-score e AUC-ROC, escolhidas por sua relevância em bases desbalanceadas. Os resultados demonstraram que o algoritmo XGBoost apresentou desempenho superior em todas as métricas, com destaque para sua capacidade discriminativa e sensibilidade. A Regressão Logística apresentou desempenho mais modesto, e o modelo de Decision Tree obteve acurácia elevada, mas com baixa capacidade de identificar casos positivos. Esses achados reforçam o potencial dos algoritmos de Machine Learning para auxiliar na tomada de decisão no mercado de crédito, contribuindo para a redução de riscos e aumento da eficiência operacional. A pesquisa também aponta direções para estudos futuros, como a ampliação do conjunto de algoritmos testados e a aplicação de técnicas de explicabilidade.

Linha temática: Contabilidade Financeira e Finanças

Palavras-Chave: Análise de crédito; Inadimplência; Aprendizado de Máquina; Classificação; Modelos Preditivos.



1. Introdução

A concessão de crédito é um dos pilares do funcionamento do sistema econômico, pois possibilita o financiamento de atividades, o estímulo a investimentos e o fortalecimento do consumo. Esse processo contribui para a alocação eficiente de recursos e para o desenvolvimento econômico, mas envolve riscos relacionados à inadimplência, que podem comprometer a estabilidade financeira de instituições e do próprio mercado (Assaf Neto, 2016). Por isso, torna-se essencial adotar métodos que assegurem maior precisão na análise de risco de crédito, auxiliando na tomada de decisão.

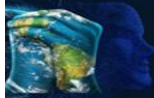
Tradicionalmente, modelos estatísticos como a regressão logística e a análise discriminante têm sido utilizados para prever inadimplência, baseando-se em variáveis como histórico financeiro, renda e situação patrimonial do solicitante. Apesar de sua ampla utilização, essas técnicas apresentam limitações, como a suposição de linearidade entre variáveis e baixa adaptabilidade a bases complexas ou desbalanceadas (Louzada et al., 2016). Tais restrições indicam que, diante da crescente complexidade dos dados, esses métodos podem não ser suficientes para capturar padrões relevantes do comportamento de crédito.

Nos últimos anos, a digitalização do sistema financeiro intensificada no Brasil por iniciativas como o *Open Finance*, o crédito digital e o PIX ampliaram a demanda por crédito e gerou volumes massivos de dados. Esse novo cenário aumenta a necessidade de métodos mais robustos e ágeis de análise, capazes de lidar com grandes bases, detectar padrões complexos e oferecer previsões mais dinâmicas. Assim, os modelos estatísticos tradicionais tornam-se desafiados, abrindo espaço para o uso de técnicas de aprendizado de máquina, *Machine Learning*, que vêm sendo aplicadas em diferentes setores do mercado financeiro.

Nesse contexto, algoritmos de ML como *XGBoost*, *Random Forest* e *CatBoost* destacam-se pela capacidade de processar múltiplas variáveis, identificar relações não lineares e fornecer resultados mais acurados (Chen & Guestrin, 2016; James et al., 2021). Além da precisão, esses modelos podem reduzir perdas financeiras ao melhorar a detecção de inadimplência e permitir decisões de concessão de crédito mais fundamentadas. Entretanto, a implementação dessas ferramentas também apresenta desafios, entre eles a interpretabilidade dos modelos complexos frequentemente caracterizados como “caixas-pretas” e as exigências de transparência e conformidade regulatória previstas na Lei Geral de Proteção de Dados (LGPD).

Diante desse cenário, este estudo busca responder à seguinte questão: **como as técnicas de *Machine Learning* podem melhorar a precisão e a eficiência na análise de risco de crédito em comparação com os métodos tradicionais?** Para tanto, o objetivo é analisar a performance de diferentes algoritmos de ML na previsão de inadimplência em operações de crédito, utilizando dados da *LendingClub*, e comparar seus resultados com os da regressão logística. Espera-se contribuir tanto para a literatura acadêmica, ao oferecer evidências sobre a eficácia dos modelos de ML, quanto para a prática profissional, ao indicar caminhos para uma análise de crédito mais precisa, eficiente e alinhada às demandas do mercado financeiro contemporâneo.

Este estudo se justifica pela necessidade de avaliar empiricamente o desempenho de modelos de *Machine Learning* na previsão do risco de inadimplência, comparando-os com métodos tradicionais a partir de dados reais da *LendingClub*. A relevância reside na urgência de se obter evidências sobre a viabilidade do ML para antecipar a inadimplência com alta precisão e agilidade, um fator crucial na era do *Open Finance*. Ao realizar essa análise comparativa (ML vs. Regressão Logística), o trabalho busca preencher uma lacuna prática e acadêmica, fornecendo evidências sobre a eficácia preditiva dessas técnicas e contribuindo para a modernização da gestão de risco no setor financeiro. Em última instância, o estudo colabora



com a construção de um sistema de crédito mais eficiente, preciso e resiliente, alinhado às exigências tecnológicas e regulatórias do mercado contemporâneo.

2. Revisão da Literatura

2.1 Fundamentos da Análise de Crédito

A análise de crédito é um dos pilares do sistema financeiro, sendo responsável por avaliar a capacidade de um indivíduo ou empresa de cumprir compromissos financeiros previamente assumidos. Para Assaf Neto (2016), o crédito envolve uma relação de confiança entre as partes, cuja manutenção depende da existência de mecanismos robustos de avaliação de risco capazes de reduzir perdas e sustentar a estabilidade do mercado. Nesse sentido, o risco de crédito, definido pelo Comitê de Supervisão Bancária de Basileia (BIS, 2001) como a possibilidade de perdas associadas ao não cumprimento das obrigações financeiras de um devedor, é inerente a todas as operações de concessão de crédito e demanda constante monitoramento.

Tradicionalmente, a mensuração desse risco foi realizada por meio de modelos estatísticos clássicos, como a regressão logística e a análise discriminante linear. Esses modelos utilizam informações socioeconômicas e financeiras renda, tempo de emprego, histórico de crédito, idade e inadimplência anterior para estimar a probabilidade de não pagamento. Sua ampla adoção ao longo das últimas décadas deve-se à simplicidade de aplicação, à interpretabilidade dos resultados e à consistência estatística alcançada (Thomas, Crook & Edelman, 2017). No entanto, embora funcionais, esses métodos apresentam limitações quanto à capacidade de capturar relações complexas entre variáveis e de se adaptar a cenários de maior incerteza.

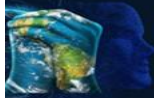
De acordo com Venkatesh e Anuradha (2019), a performance dos modelos depende diretamente da qualidade e relevância das variáveis utilizadas. Além de fatores objetivos, como renda e patrimônio, variáveis de natureza comportamental e contextos macroeconômicos exercem influência significativa na inadimplência. Essa complexidade sugere que métodos baseados exclusivamente em técnicas estatísticas podem não ser suficientes para lidar com a crescente heterogeneidade dos tomadores de crédito, especialmente em ambientes de instabilidade econômica ou em mercados caracterizados por rápidas mudanças tecnológicas.

Assim, a análise de crédito deve ser entendida como uma prática que se situa na interseção entre estatística, economia comportamental e regulamentação bancária. A evolução do setor financeiro e a ampliação da disponibilidade de dados digitais vêm demandando abordagens mais flexíveis e precisas, capazes de superar as limitações dos modelos tradicionais. Essa base teórica abre espaço para a discussão das técnicas contemporâneas, em especial os métodos de aprendizado de máquina, que buscam aprimorar a previsão de inadimplência e apoiar a modernização da gestão do risco de crédito.

2.2 Modelos Tradicionais de Análise de Crédito

A análise de crédito é historicamente sustentada por modelos estatísticos que visam quantificar o risco de inadimplência a partir de características observáveis dos solicitantes de crédito. Esses modelos, conhecidos como modelos tradicionais de *credit scoring*, foram amplamente difundidos entre as instituições financeiras a partir da década de 1960.

A análise discriminante, conforme explicam Press et al. (2016), é uma técnica estatística multivariada que busca encontrar uma função linear capaz de separar dois grupos por exemplo, clientes inadimplentes e adimplentes a partir de variáveis explicativas como renda, idade, histórico de crédito e tempo de emprego. A regressão logística, por sua vez, é um modelo probabilístico que estima a chance de ocorrência de um evento binário (como o *default* ou



inadimplência), sendo amplamente utilizada devido à sua robustez, simplicidade de interpretação e aderência a dados categóricos e contínuos.

Ambos os métodos compartilham características importantes: são interpretáveis, apresentam baixo custo computacional e têm sido, por décadas, a espinha dorsal das decisões de crédito em bancos, *fintechs* e empresas de análise de risco. No Brasil, a adoção desses modelos é endossada por regulamentações do Banco Central e pelas práticas adotadas por birôs de crédito (como Serasa *Experian* e Boa Vista), que frequentemente operam com scores baseados nesses métodos (Assaf Neto, 2016).

Apesar de sua ampla utilização, diversos estudos têm apontado limitações desses modelos no cenário contemporâneo. Em especial, a rigidez matemática imposta por essas técnicas, que assumem linearidade nas relações entre variáveis e independência entre os atributos, tende a reduzir sua eficácia quando aplicadas a bases de dados complexas e não estruturadas. Como destacam Syazwan et al. (2021), essas limitações resultam em baixa acurácia preditiva em cenários nos quais os dados são desbalanceados (com número desproporcional de inadimplentes/adimplentes) ou onde há múltiplas interações não lineares entre variáveis.

Além disso, a adaptabilidade desses modelos a contextos mutáveis, como crises econômicas, mudanças regulatórias ou novas práticas digitais de consumo, é restrita. Por exemplo, mudanças no perfil do consumidor digital, como comportamento em redes sociais ou uso de serviços de pagamento instantâneo, dificilmente são capturadas com precisão por modelos tradicionais. Nesse sentido, a literatura recente tem discutido a aplicabilidade desses métodos no atual ambiente financeiro dinâmico (Yurdakul, 2014).

2.3 Machine Learning Aplicado à Análise de Crédito

2.3.1 Conceitos e Funcionamento do Machine Learning

O *Machine Learning* (ML), ou aprendizado de máquina, é um subcampo da inteligência artificial que busca desenvolver algoritmos capazes de aprender padrões a partir de dados e realizar previsões ou classificações sem serem explicitamente programados para isso. Segundo Géron (2022), o principal diferencial do ML está na sua capacidade de adaptar-se a grandes volumes de dados, modelar relações complexas e gerar resultados preditivos com alta acurácia, tornando-se especialmente útil em áreas que envolvem grande variabilidade de cenários, como a análise de crédito.

Em contextos financeiros, a aplicação de ML é predominante em tarefas como detecção de fraudes, segmentação de clientes, precificação de produtos e, de forma destacada, na avaliação de risco de crédito. Nessa área, os algoritmos são treinados com dados históricos de concessão de crédito (com ou sem inadimplência) para aprender padrões que indiquem maior ou menor probabilidade de default. Esse processo é conhecido como aprendizado supervisionado, no qual os modelos são alimentados com dados rotulados, ou seja, os exemplos incluem a variável-alvo de interesse, como “inadimplente” ou “adimplente”.

Entre os principais tipos de aprendizado, destacam-se: o Aprendizado supervisionado que se trata do modelo que aprende a partir de pares de entrada e saída já conhecidos e o Aprendizado não supervisionado que é o modelo que não recebe rótulos nas saídas, sendo utilizado para identificar agrupamentos ou padrões ocultos nos dados. É útil, por exemplo, para segmentação de perfis de clientes com base em comportamento financeiro.

A adoção de ML na análise de crédito é motivada pela sua eficiência em lidar com bases de dados extensas e heterogêneas, compostas por variáveis de natureza diversa (numéricas, categóricas, temporais), bem como pela sua capacidade de capturar relações não lineares e de reconhecer interações complexas entre variáveis. Isso amplia significativamente a acurácia das



previsões, especialmente em comparação com modelos tradicionais, que exigem suposições mais restritivas e têm menor flexibilidade (Chang et al., 2018).

Além disso, modelos de ML oferecem recursos como validação cruzada, ajuste de hiperparâmetros e regularização, que aumentam a confiabilidade das estimativas geradas. Tais atributos fazem com que as técnicas de aprendizado de máquina sejam cada vez mais adotadas no setor bancário e pelas *fintechs*, que buscam decisões de crédito mais rápidas, precisas e escaláveis (Nonato, 2022).

2.3.2 Algoritmos de Machine Learning

A aplicação de técnicas de ML na análise de risco de crédito tem se consolidado em função do desempenho superior de determinados algoritmos em tarefas de classificação binária, como prever se um cliente será ou não inadimplente. Dentre os algoritmos mais utilizados e pesquisados na literatura, destacam-se os modelos de árvore de decisão, como o *Decision Tree*, e suas variantes em *ensemble learning*, como o *XGBoost*.

O *Decision Tree* (árvore de decisão) é um algoritmo de aprendizado supervisionado que organiza as decisões por meio de uma estrutura hierárquica em forma de árvore, em que cada nó interno representa uma condição de divisão com base em uma variável explicativa. Segundo Géron (2022), esse modelo é amplamente utilizado por sua intuitividade, facilidade de visualização e boa capacidade de interpretação, especialmente em contextos que exigem explicabilidade dos resultados. Embora possa apresentar alta variabilidade em classificações isoladas, sua estrutura permite capturar padrões não lineares nos dados, sendo útil em tarefas de classificação de risco de crédito, mesmo com variáveis de diferentes naturezas.

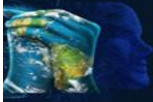
O *XGBoost* (Extreme Gradient Boosting) é uma técnica de *boosting* que gera modelos sequenciais, em que cada novo modelo corrige os erros cometidos pelos anteriores. Chen e Guestrin (2016) desenvolveram o algoritmo com foco em alta performance, regularização contra *overfitting* e velocidade computacional, características que o tornaram referência em competições de ciência de dados e em aplicações de crédito em larga escala. Chang et al. (2018) demonstraram que o *XGBoost* apresentou o melhor desempenho entre vários algoritmos analisados, atingindo 99,4% de acurácia em previsões de inadimplência.

A escolha do algoritmo mais adequado depende de uma série de fatores, incluindo a natureza dos dados, o objetivo da aplicação (precisão x interpretabilidade), o volume de dados disponíveis e as exigências regulatórias. Assim, a literatura recente tem apontado os algoritmos baseados em árvores de decisão como primeiras opções, em razão de seu desempenho consistente, adaptabilidade e razoável transparência nos resultados.

2.3.3 Vantagens do Machine Learning para a Análise de Crédito

A adoção de algoritmos de ML na análise de crédito tem ganhado destaque devido às vantagens técnicas, operacionais e estratégicas que essas tecnologias oferecem em relação aos métodos tradicionais. Tais benefícios se refletem diretamente em melhorias no processo decisório, na redução de perdas por inadimplência e na otimização da concessão de crédito. Uma das principais vantagens está na capacidade de lidar com grandes volumes de dados heterogêneos, algo comum em instituições financeiras que trabalham com variáveis cadastrais, transacionais e comportamentais. Os algoritmos de ML são capazes de processar, em alta velocidade, dados estruturados e não estruturados, como informações textuais, categorias nominais, séries temporais e históricos de crédito (Izbicki, 2022).

Outro benefício relevante é a capacidade de identificar padrões complexos e relações não lineares entre variáveis. Enquanto modelos tradicionais, como a regressão logística, assumem uma relação linear entre os atributos e a variável-alvo, os modelos de ML podem capturar interações sutis e até não evidentes entre características do cliente, aumentando a



precisão preditiva (Izbicki, 2022). Lai et al. (2021) destacam que esse tipo de modelagem contribui para a redução de falsos positivos e negativos, especialmente em bases desbalanceadas, comuns em cenários de crédito. Além disso, técnicas como o *ensemble learning* (como *Random Forest*, *XGBoost* e *CatBoost*) agregam robustez ao modelo ao combinar múltiplos preditores, resultando em maior estabilidade dos resultados mesmo diante de pequenas variações nos dados (Dong et al., 2020).

Essa característica é especialmente útil em contextos econômicos voláteis, nos quais a inadimplência pode sofrer influência de fatores macroeconômicos externos. O ML também proporciona ganhos operacionais, permitindo a automatização da tomada de decisão em concessões de crédito com alta acurácia e em tempo real. Em plataformas digitais, isso se traduz em processos mais ágeis, escaláveis e integráveis a sistemas de avaliação contínua de risco, otimizando a experiência do cliente e reduzindo custos de análise.

2.3.4 Outros estudos sobre Machine Learning na análise de crédito

O uso de técnicas de ML para análise de risco de crédito tem se intensificado nos últimos anos, impulsionado pela crescente disponibilidade de dados e pela necessidade de previsões mais precisas em ambientes financeiros cada vez mais complexos. Estudos empíricos recentes, tanto no cenário nacional quanto internacional, reforçam a superioridade dos algoritmos de ML em comparação com os métodos estatísticos tradicionais em termos de desempenho preditivo e adaptabilidade.

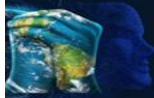
Coelho, Amorim e Camargos (2021) conduziram um estudo comparativo no qual *Random Forest*, *XGBoost* e regressão logística clássica foram aplicados a um conjunto de dados de crédito. Os resultados indicaram que os métodos de ML apresentaram capacidade preditiva melhor quando comparados com a regressão logística. O *XGBoost* obteve o melhor desempenho, evidenciando a eficácia dos modelos de *ensemble learning* em capturar padrões complexos e não lineares em dados financeiros.

Silva (2022) investigou a aplicação dos algoritmos de *machine learning* *Random Forest* e *Gradient Boosting* e da técnica de regressão logística para identificação de inadimplência de contratos habitacionais. A pesquisa apresentou resultados bem próximos entre todos os modelos testados, mas concluiu que o modelo *Gradient Boosting* apresentou as melhores métricas em termos de acurácia e eficiência quanto à previsão e sensibilidade.

No estudo de Pereira (2020), foram avaliados os algoritmos KNN, SVM, *Decision Tree* e *Random Forest* em tarefas de previsão de inadimplência. O autor constatou que o *Random Forest* apresentou o melhor desempenho geral em termos de acurácia e robustez contra *overfitting*, reforçando a tendência observada em outras pesquisas que apontam essa técnica como uma das mais eficazes em cenários de classificação binária.

Em âmbito internacional, Melsom et al. (2022) publicaram um estudo que utilizou modelos de machine learning para identificar inadimplência de crédito em dados do *Norwegian Bank*. Os autores compararam os resultados do modelo *Light Gradient Boosting Machine* (LightGBM) com regressão logística e obtiveram resultados consistentemente melhores com o modelo de *machine learning*. Além disso, os autores testaram métodos de explicabilidade, como o SHAP, para entender mais a fundo a influência de cada variável.

Chang (2024) realizou um estudo comparativo aplicando Redes Neurais, Regressão Logística, *AdaBoost*, *XGBoost*, e LightGBM para prever inadimplência em crédito pessoal, utilizando métricas como AUC, precisão e recall. O *XGBoost* apresentou a melhor performance global, alcançando acurácia de 99.4%, sendo recomendado como o modelo preferencial para aplicações em risco de crédito devido à sua alta capacidade de generalização e baixo erro de classificação.



Em síntese, a literatura recente confirma a superioridade dos modelos de ML em aplicações de análise de crédito, particularmente em cenários que exigem o tratamento de bases extensas, desbalanceadas e com múltiplas variáveis interdependentes. No entanto, observa-se uma carência de estudos que apliquem essas técnicas especificamente em bases públicas, como o *LendingClub*, sob diferentes estratégias de seleção de variáveis e hiperparâmetros. Essa lacuna reforça a relevância do presente estudo.

3. Materiais e Métodos

Este trabalho caracteriza-se como uma pesquisa aplicada, pois visa resolver um problema prático relacionado à análise e previsão de inadimplência no contexto financeiro, por meio da utilização de técnicas de aprendizado de máquina (*Machine Learning*). Trata-se, ainda, de uma pesquisa de abordagem quantitativa, uma vez que se baseia em dados numéricos, métricas estatísticas e modelos matemáticos para gerar conclusões objetivas.

Quanto aos objetivos, a pesquisa é classificada como descritiva e preditiva. É descritiva porque busca compreender o comportamento e a distribuição das variáveis relacionadas à concessão de crédito e inadimplência dentro da base de dados utilizada, mapeando padrões relevantes; e é preditiva porque seu foco principal está na avaliação do desempenho de diferentes algoritmos de classificação binária, com o intuito de prever a probabilidade de inadimplência a partir de variáveis de perfil financeiro e comportamental dos clientes.

Por fim, em relação aos procedimentos técnicos, este trabalho configura-se como uma pesquisa experimental computacional, pois envolve a construção, teste e comparação de diferentes modelos algorítmicos sob um mesmo conjunto de dados, buscando avaliar seu desempenho de forma objetiva com base em métricas pré-estabelecidas.

3.1 Amostra e Fonte de Dados

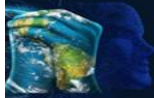
Para a realização deste estudo, foi utilizada uma base de dados obtida da plataforma digital Kaggle, referente a empréstimos concedidos pela plataforma LendingClub, uma *fintech* norte-americana especializada na concessão de crédito pessoal online, contendo originalmente aproximadamente 2.2 milhões de registros e 151 variáveis. A escolha dessa base justifica-se por sua ampla adoção em estudos acadêmicos e práticos na área de risco de crédito, além de apresentar um conjunto robusto de dados históricos com diversas variáveis financeiras e comportamentais dos solicitantes de empréstimos.

O período de abrangência dos dados vai de 2007 a 2018, cobrindo diferentes ciclos econômicos e permitindo a análise de comportamentos variados ao longo do tempo. A variável-alvo da modelagem preditiva é a inadimplência, nomeada na base de dados como “loan_status” e representada inicialmente por rótulos de empréstimo completamente pago (Fully Paid), empréstimo em andamento (Current) e empréstimo não pago (Charged Off), além de outras classificações minoritárias relativas a faixas de rolagens nos atrasos.

Como o objetivo deste trabalho concentrou-se nos empréstimos completamente pagos ou não pagos, a base de dados foi filtrada apenas com essas categorias, restando 1.345.310 registros. Além disso, para fins de classificação binária, a variável foi convertida em um formato numérico com duas categorias: 0 para adimplente e 1 para inadimplente. Essa transformação viabiliza a aplicação de algoritmos de aprendizado supervisionado.

As variáveis independentes da base incluem atributos como pontuação de crédito, renda anual, relação dívida/renda, finalidade do empréstimo, entre outras. Cada uma dessas variáveis possui potencial explicativo sobre o comportamento financeiro do cliente e sua propensão à inadimplência.

Um aspecto importante da base é o desbalanceamento de classes, ou seja, a proporção entre clientes inadimplentes e adimplentes não é equilibrada. Em geral, a maioria dos registros



representa clientes que pagaram seus empréstimos normalmente. Essa característica reforça a necessidade de utilizar métricas de avaliação apropriadas e algoritmos capazes de lidar com a detecção de eventos raros, como é o caso da inadimplência.

3.2 Seleção de Variáveis

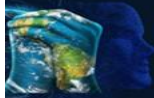
A seleção de variáveis foi conduzida para assegurar a qualidade dos dados e evitar vieses que comprometessem a validade dos resultados, como redundâncias, granularidades excessivas ou vazamento de dados (data leakage). Essa etapa é essencial no pipeline de aprendizado de máquina, pois atributos irrelevantes ou inconsistentes podem prejudicar tanto a performance quanto a interpretabilidade dos modelos (Kuhn et al., 2013).

Inicialmente, foram excluídas variáveis identificadoras e não padronizadas, por apresentarem alta variabilidade e baixo potencial de generalização. Em seguida, adotou-se como critério a exclusão de variáveis com mais de 30% de valores ausentes, conforme recomendações de Acuña e Rodríguez (2004), o que resultou em 87 variáveis consideradas suficientemente completas para análise.

Na sequência, procedeu-se à remoção de atributos que poderiam causar data leakage, isto é, variáveis que carregavam informações posteriores à concessão do crédito ou à definição da variável-alvo, o que comprometeria a validade preditiva (Kapoor & Narayanan, 2023). Após essa filtragem, 29 variáveis foram selecionadas para as etapas de pré-processamento e modelagem, apresentadas no Quadro 1.

Quadro 1: Descrição das variáveis restantes

Variável	Descrição
loan_amnt	Valor do empréstimo solicitado pelo cliente.
funded_amnt	Valor efetivamente liberado pela instituição financeira.
Term	Prazo total do empréstimo (por exemplo, 36 ou 60 meses).
int_rate	Taxa de juros anual aplicada ao empréstimo.
Installment	Valor da parcela mensal que o cliente deve pagar.
Grade	Classificação de risco atribuída internamente pela instituição (de A a G).
sub_grade	Subclassificação da nota de risco (por exemplo, A1, A2, etc.).
emp_length	Tempo de emprego declarado pelo cliente (em anos).
home_ownership	Situação de moradia do cliente (própria, alugada, hipotecada, etc.).
annual_inc	Renda anual declarada pelo cliente.
verification_status	Se a renda ou informações financeiras foram verificadas.
Purpose	Finalidade do empréstimo (ex: carro, casa, dívida).
Dti	Proporção entre dívida e renda do cliente.
delinq_2yrs	Número de ocorrências de atraso nos últimos dois anos.



inq_last_6mths	Consultas ao histórico de crédito nos últimos seis meses.
open_acc	Número de contas de crédito atualmente em aberto.
pub_rec	Registros públicos negativos (ex: falências, execuções).
revol_bal	Saldo atual em contas rotativas (ex: cartão de crédito).
revol_util	Percentual de utilização do limite rotativo de crédito.
total_acc	Total de contas de crédito que o cliente teve ao longo da vida.
initial_list_status	Status de listagem do empréstimo (individual ou em grupo).
application_type	Tipo de aplicação (individual ou conjunta).
acc_now_delinq	Número de contas atualmente inadimplentes.
tot_coll_amt	Valor total em cobranças registradas.
tot_cur_bal	Saldo total atual em todas as contas do cliente.
total_rev_hi_lim	Limite total de crédito rotativo disponível.
acc_open_past_24mths	Número de contas abertas nos últimos 24 meses.
avg_cur_bal	Saldo médio atual nas contas do cliente.
mort_acc	Número de contas relacionadas a hipoteca ou financiamento imobiliário.

Fonte: Elaborado pelo autor

Essas 29 variáveis finais foram consideradas adequadas tanto do ponto de vista estatístico quanto conceitual, pois representam atributos disponíveis no momento da análise de crédito e possuem relação potencial com o comportamento de pagamento do cliente.

3.3 Pré-processamento dos Dados

O pré-processamento constitui etapa essencial em projetos de aprendizado de máquina, pois a qualidade dessa preparação impacta diretamente a performance, a capacidade de generalização e a confiabilidade dos modelos (Géron, 2022).

Primeiramente, as variáveis categóricas foram transformadas por meio da técnica one-hot encoding, que converte categorias em colunas binárias, evitando interpretações incorretas de ordem entre classes e preservando a granularidade necessária para a modelagem (Bruce, Bruce & Gedeck, 2020). Como resultado, a base expandiu-se de 29 variáveis para 71 colunas, refletindo o acréscimo decorrente da codificação.

Em seguida, realizou-se a análise de dados ausentes, considerando que valores faltantes podem gerar vieses e perda de eficiência estatística (Bruce, Bruce & Gedeck, 2020). Foram identificados seis variáveis com nulos, variando entre 0,03% e 7,85% do total de registros, conforme apresentado na Tabela 1. Esse diagnóstico permitiu delinear estratégias adequadas para tratamento, assegurando a integridade da base antes da etapa de modelagem.



Tabela 1: Variáveis com dados ausentes

Variável	Quantidade de Nulos	% do total
mo_sin_old_il_acct	105.575	7,85%
mo_sin_old_rev_tl_op	67.528	5,02%
mort_acc	47.281	3,51%
revol_util	857	0,06%
pub_rec_bankruptcies	697	0,05%
Dti	374	0,03%

Fonte: Elaborado pelo autor

Os modelos deste estudo utilizaram aprendizado supervisionado, apropriado para problemas de classificação binária, como a previsão de inadimplência, em que a variável-alvo é conhecida durante o treinamento (Kelleher; Namee; D'Arcy, 2015). Para avaliar o desempenho de forma realista e evitar overfitting, os dados foram divididos em dois subconjuntos: treinamento e teste. Essa separação permite que o modelo seja testado em dados não utilizados no treinamento, garantindo sua capacidade de generalização e possibilitando a comparação objetiva entre as previsões e os valores reais da variável-alvo (Géron, 2022; Ying, 2019).

3.4 Métricas de avaliação

Para avaliar de forma adequada o desempenho dos modelos de classificação utilizados neste estudo, foi necessário selecionar métricas que não apenas reflitam a proporção de acertos, mas também capturem a capacidade do modelo de distinguir corretamente entre as classes, especialmente em bases desbalanceadas como é comum em estudos de inadimplência. Conforme Géron (2022), a forma de medir o desempenho influencia diretamente decisões sobre qual modelo adotar, ajustar ou descartar. Avaliar métricas isoladamente pode gerar interpretações enganosas, pois um modelo pode apresentar alta acurácia simplesmente por classificar corretamente a classe majoritária, sem identificar adequadamente a classe minoritária. Por esse motivo, adotou-se um conjunto de métricas complementares que permite uma análise mais completa da performance do modelo (Naidu, Zuva & Sibanda, 2023). A Tabela 2 apresenta essas métricas, incluindo sua fórmula, objetivo e fundamentação teórica.

Tabela 2 – Métricas de avaliação utilizadas no estudo

Métrica	Fórmula	Objetivo	Autores
Acurácia	$Acurácia = \frac{VP + VN}{VP + VN + FP + FN}$	Mede a proporção de classificações corretas (positivas e negativas) em relação ao total de observações. Fornece visão geral da taxa de acerto, mas pode ser limitada em bases desbalanceadas.	Géron (2022)
Recall (Sensibilidade)	$Recall = \frac{VP}{VP + FN}$	Avalia a capacidade do modelo de identificar corretamente os casos positivos (inadimplentes). Penaliza falsos negativos.	Strauss, Júnior e Ferreira (2022)
F1-Score	$F1 - Score = 2 \times \frac{Precisao \times Recall}{Precisao + Recall}$	Combina precisão e recall em uma média harmônica, equilibrando custos de falsos positivos e falsos negativos. Indicado para bases desbalanceadas.	Strauss, Júnior e Ferreira (2022)
AUC-ROC	Área sob a curva ROC (TPR × FPR)	Mede a capacidade do modelo de distinguir entre classes positivas e negativas, independentemente do	Narkhede (2018)



Métrica	Fórmula	Objetivo	Autores
		limiar de decisão. Quanto mais próximo de 1, melhor o poder discriminativo.	

Fonte: Elaborado pelo autor

4. Resultados

4.1 Avaliação do modelo Regressão Logística

A regressão logística foi utilizada como modelo de referência (baseline) neste trabalho por ser uma técnica amplamente adotada na classificação binária de eventos. Essa escolha se deve ao fato de se tratar de um modelo amplamente difundido e explorado, além de apresentar simplicidade de implementação e interpretação. Ao ser aplicada à base de dados deste estudo, a regressão logística apresentou os seguintes resultados:

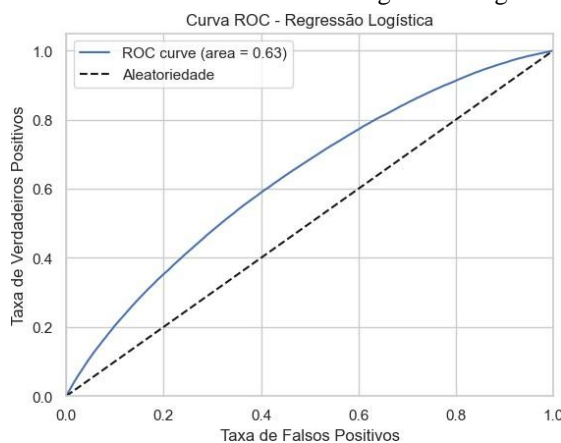
Tabela 3: Métricas alcançadas pela Regressão Logística

Métrica	Valor
AUC-ROC	0,633
Acurácia	0,599
Recall	0,588
F1-Score	0,371

Fonte: Elaborado pelo autor

O AUC-ROC de 0,633 indica uma capacidade discriminativa abaixo do desejável para modelos de risco de crédito. Na prática, isso significa que o modelo apresenta dificuldade em separar corretamente os casos de inadimplência e adimplência.

Gráfico 1: AUC-ROC Regressão Logística



Fonte: Elaborado pelo autor

Já a acurácia de 59,9% é apenas marginalmente superior à de um modelo aleatório, o que também reforça sua limitação nesse contexto.

A sensibilidade (recall) de 0,588 revela que o modelo consegue identificar cerca de 59% dos clientes inadimplentes. Embora seja um resultado modesto, é importante destacar que essa métrica é mais relevante do que a acurácia quando se trabalha com bases desbalanceadas, como é o caso neste estudo. O F1-score de 0,371, por sua vez, evidencia um desequilíbrio entre a capacidade de acerto (precisão) e a sensibilidade (recall), o que compromete a performance geral do modelo.

Esses resultados estão alinhados com o que se observa na literatura. Conforme Silva (2022), a regressão logística tende a apresentar desempenho insatisfatório em problemas com



forte assimetria de classes e com interações não lineares entre variáveis explicativas, como ocorre na análise de inadimplência.

Portanto, embora seja uma ferramenta válida como benchmark, a regressão logística demonstrou limitações importantes no presente estudo, o que reforça a necessidade de modelos mais robustos para capturar os padrões de risco presentes nos dados de crédito.

4.2 Avaliação do algoritmo *Decision Tree*

Após o treinamento e teste do modelo gerado pelo algoritmo de *Decision Tree*, foram obtidas as seguintes métricas:

Tabela 4: Métricas alcançadas pelo algoritmo *Decision Tree*

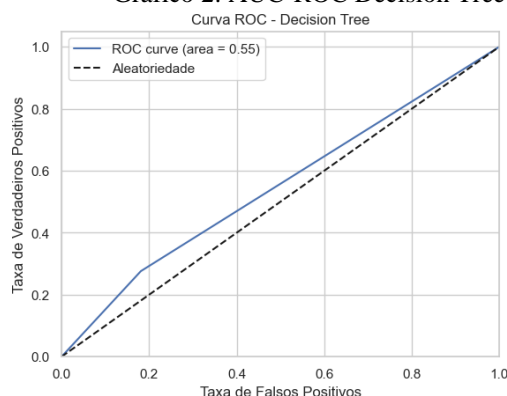
Métrica	Valor
AUC-ROC	0,549
Acurácia	0,709
Recall	0,283
F1-Score	0,281

Fonte: Elaborado pelo autor

A acurácia de 70,9% pode sugerir um bom desempenho geral, mas trata-se de um efeito ilusório causado pelo desbalanceamento da base de dados, uma vez que o modelo tende a classificar a maioria das instâncias como adimplentes. Isso é evidenciado pelo recall de apenas 28,3%, o que indica que o modelo falha em reconhecer corretamente a maior parte dos inadimplentes a classe de maior interesse neste estudo.

Adicionalmente, o F1-score de 0,281 demonstra um fraco equilíbrio entre precisão e sensibilidade, sugerindo que o modelo apresenta elevado número de erros ao tentar prever a classe positiva. A AUC-ROC de 0,549 reforça essa limitação, evidenciando que a árvore de decisão possui baixa capacidade de distinguir entre as duas classes seu desempenho está muito próximo de um modelo aleatório (AUC = 0,5).

Gráfico 2: AUC-ROC Decision Tree



Fonte: Elaborado pelo autor

Esses resultados indicam que, embora a *Decision Tree* apresente alguma capacidade de classificação, seu uso isolado não é eficaz na detecção de inadimplência em contextos com variáveis complexas e dados desbalanceados. Modelos mais avançados, como os baseados em *ensemble*, tendem a contornar essas limitações ao combinar múltiplas árvores com ajustes de pesos e penalizações.



4.3 Avaliação do algoritmo XGBoost

Da mesma forma, o algoritmo XGBoost também foi treinado e apresentou as métricas de desempenho abaixo.

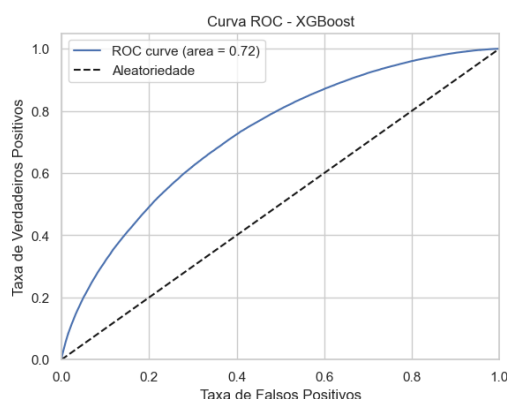
Tabela 5: Métricas alcançadas pelo algoritmo XGBoost

Métrica	Valor
AUC-ROC	0,726
Acurácia	0,653
Recall	0,682
F1-Score	0,441

Fonte: Elaborado pelo autor

O AUC-ROC de 0,726 indica uma boa capacidade de discriminação entre inadimplentes e adimplentes, mostrando que o XGBoost consegue, de fato, separar os dois grupos com um nível razoável de confiança. É o valor mais alto obtido entre os modelos testados até o momento, o que já demonstra sua superioridade nesse aspecto.

Gráfico 3: AUC-ROC XGBoost



Fonte: Elaborado pelo autor

O recall de 68,2% mostra que o modelo tem uma boa sensibilidade, conseguindo identificar corretamente a maioria dos inadimplentes — um fator crítico quando o objetivo é mitigar risco financeiro. O F1-score de 0,441, embora ainda moderado, demonstra um equilíbrio superior entre precisão e recall em relação aos demais modelos, sendo o mais alto dos três testados.

Já a acurácia de 65,3% representa um avanço considerável frente à regressão logística (59,9%), mesmo não sendo o foco principal da avaliação, dada a natureza desbalanceada dos dados. Esses resultados reforçam a efetividade do XGBoost em problemas de análise de crédito, alinhando-se com os estudos de Wang et al. (2018), que destacam seu bom desempenho mesmo sem ajustes extremos de hiperparâmetros, e com Lopes, Colombi e Mutz (2023), que também demonstraram que modelos mais avançados, como o XGBoost, superam significativamente a regressão logística nesse tipo de tarefa.

4.4 Discussão dos resultados

A partir dos resultados obtidos com os modelos de Regressão Logística, Decision Tree e XGBoost, é possível realizar uma análise comparativa que evidencia não apenas as diferenças de desempenho numérico, mas também as implicações práticas de suas limitações e potenciais



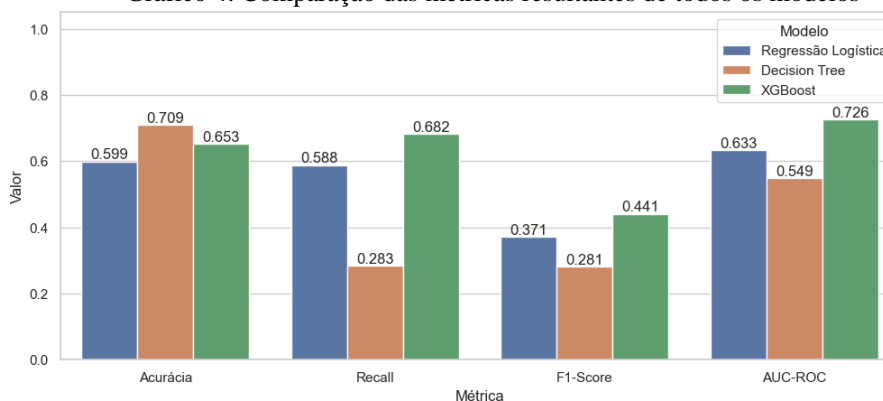
na previsão de inadimplência. A Tabela 6 e o Gráfico 4 resumem os principais indicadores utilizados para avaliar o desempenho dos modelos.

Tabela 6: Comparação das métricas resultantes de todos os modelos

Modelo	Acurácia	Recall	F1-Score	AUC-ROC
Regressão Logística	0,599	0,588	0,371	0,633
Decision Tree	0,709	0,283	0,281	0,549
XGBoost	0,653	0,682	0,441	0,726

Fonte: Elaborado pelo autor

Gráfico 4: Comparação das métricas resultantes de todos os modelos

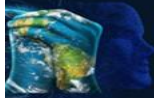


Fonte: Elaborado pelo autor

Observa-se que o modelo *Decision Tree* apresentou a maior acurácia (70,9%), o que poderia sugerir, de forma superficial, um bom desempenho. No entanto, ao se analisarem métricas mais sensíveis à identificação de inadimplentes, como o recall e o F1-score, o modelo demonstrou deficiências. O recall de apenas 28,3% indica que a maioria dos inadimplentes não foi corretamente identificada, o que, em um contexto de crédito, pode representar uma exposição significativa a riscos de inadimplência.

Esse contraste entre alta acurácia e baixo recall é comum em bases desbalanceadas, como ocorre em cenários de crédito, nos quais os adimplentes representam a ampla maioria dos registros. Em algoritmos como a *Decision Tree*, as divisões são feitas para maximizar métricas de pureza, que tendem a privilegiar a classe dominante na tentativa de aumentar a acurácia global. Com isso, o modelo aprende regras que classificam corretamente a maioria, mas são pouco eficazes para detectar a minoria (os inadimplentes), que são justamente os casos mais relevantes. Isso explica o recall extremamente baixo observado, pois muitos inadimplentes foram classificados incorretamente como bons pagadores. Na prática, isso significaria aprovar crédito para quem tem alto risco de inadimplência, comprometendo a eficácia operacional do modelo. Como discutido por Louzada et al. (2016), modelos simples como árvores de decisão apresentam desempenho limitado quando submetidos a conjuntos de dados com forte desbalanceamento entre classes.

Em problemas de crédito, o uso da acurácia como principal métrica pode levar a conclusões equivocadas, especialmente em bases desbalanceadas. Isso porque a acurácia contabiliza todos os acertos, sem diferenciar o custo dos erros. Nesse contexto, o recall assume papel central, pois representa a proporção de inadimplentes corretamente identificados. Um falso negativo, ou seja, um inadimplente que o modelo classificou erroneamente como adimplente pode gerar prejuízos significativos à instituição, ao passo que um falso positivo apenas restringe o crédito de um cliente potencialmente bom. Por isso, modelos com maior recall são preferíveis, mesmo que apresentem menor acurácia global. Essa lógica é reforçada



por Strauss, Júnior e Ferreira (2022), que destacam a importância de utilizar métricas de avaliação ajustadas ao impacto real dos erros cometidos pelos modelos preditivos.

Por sua vez, a Regressão Logística apresentou resultados mais equilibrados, com recall de 58,8% e F1-score de 0,371, superando o *Decision Tree* na capacidade de detecção da classe de maior interesse. Contudo, os resultados ainda são limitados, especialmente ao se considerar o AUC-ROC de 0,633, que revela uma capacidade restrita do modelo em discriminar entre adimplentes e inadimplentes.

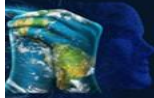
Embora a Regressão Logística não tenha apresentado o melhor desempenho entre os modelos testados, sua inclusão no estudo continua sendo relevante. Trata-se de um método consagrado, de fácil implementação e alta interpretabilidade, o que o torna valioso em contextos regulatórios e operacionais. Em ambientes que exigem justificativas claras sobre os critérios de concessão de crédito, como os moldados pela Lei Geral de Proteção de Dados (LGPD), modelos lineares como a regressão são mais facilmente auditáveis, oferecendo respostas diretas e claras sobre o impacto de cada variável. Além disso, ela funciona como um benchmark confiável, frequentemente utilizado como linha de base para comparação com algoritmos mais sofisticados. Essa função de referência, aliada à sua robustez matemática, explica sua presença contínua em aplicações reais, mesmo diante do avanço dos métodos mais complexos.

O modelo XGBoost apresentou os melhores resultados em todas as métricas analisadas, com destaque para o AUC-ROC de 0,726 e o recall de 68,2%, que indicam desempenho superior. O F1-score de 0,441 evidencia que o modelo alcançou maior equilíbrio entre precisão e sensibilidade, mostrando-se mais adequado ao problema de previsão de inadimplência, por reduzir o risco de classificações falsas negativas que implicam custos operacionais elevados sem restringir o crédito de clientes adimplentes.

O bom desempenho do XGBoost neste estudo pode ser explicado por sua arquitetura baseada em boosting, na qual diversas árvores de decisão simples são treinadas em sequência, sendo que cada nova árvore busca corrigir os erros cometidos pelas anteriores. Essa abordagem iterativa permite que o modelo aprenda padrões complexos, mesmo em bases de dados ruidosas ou desbalanceadas. Com isso, o XGBoost se mostra mais eficaz em capturar os casos de inadimplência, como evidenciado pelas suas métricas em comparação aos outros modelos testados. Estudos como os de Chen e Guestrin (2016) e Chang (2024) já demonstram que o XGBoost apresenta resultados superiores em tarefas de classificação complexas, o que reforça a consistência dos achados deste trabalho.

Ao comparar as métricas obtidas pelo modelo XGBoost neste estudo com os resultados apresentados por Chang (2024), observa-se uma diferença significativa na acurácia: 71% neste trabalho, frente a 99,4% no estudo citado. Essa diferença pode ser explicada por diversos fatores. O conjunto de dados utilizado por Chang (2024) era menos extenso e apresentava menos variáveis, além de ter sido aplicado técnicas de balanceamento das classes por meio da geração de dados sintéticos, o que favorece a performance final do modelo. Ainda assim, é importante destacar que o XGBoost também apresentou desempenho superior neste estudo, mantendo-se como o melhor algoritmo entre os testados. Essa consistência reforça os achados discutidos na Seção 2.3.4, em que evidências empíricas recentes apontam o XGBoost como um dos métodos mais eficazes para tarefas de classificação em risco de crédito.

Além do desempenho técnico, cabe destacar que modelos como o XGBoost se beneficiam de técnicas avançadas de aprendizado, como o boosting, que permite capturar padrões não lineares e interações entre variáveis, aspectos usualmente ignorados por modelos mais simples, como o *Decision Tree* ou a Regressão Logística. Isso sustenta o argumento de que, em cenários reais de análise de crédito, a adoção de algoritmos mais sofisticados não apenas melhora o desempenho estatístico, mas também contribui para decisões mais seguras e fundamentadas, favorecendo os resultados financeiros das instituições.



Como próximos passos desta pesquisa, destaca-se a possibilidade de incorporar outros algoritmos relevantes na literatura, como o Random Forest e o CatBoost, que não puderam ser testados neste estudo devido às limitações computacionais momentâneas. A inclusão desses modelos poderá enriquecer a análise comparativa e oferecer novos insights sobre o comportamento de classificadores em bases desbalanceadas. Além disso, recomenda-se a aplicação de métodos de interpretabilidade, como o SHAP (*SHapley Additive exPlanations*), que permitem identificar a contribuição de cada variável para as decisões do modelo. A explicabilidade vem ganhando espaço nas aplicações de inteligência artificial, especialmente em domínios sensíveis como o crédito, nos quais a rastreabilidade e a transparência das decisões são essenciais. Conforme discutido por Belle e Papantonis (2021), ferramentas de interpretação aumentam a confiança no uso de modelos de IA e ajudam a alinhar práticas algorítmicas com princípios éticos e regulatórios.

5. Conclusão

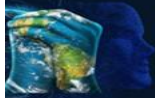
Este trabalho teve como objetivo central investigar a aplicabilidade de técnicas de aprendizado de máquina na previsão de inadimplência em operações de crédito, comparando o desempenho de diferentes algoritmos com o modelo estatístico clássico de regressão logística. A proposta fundamentou-se na necessidade crescente por métodos preditivos mais robustos, capazes de lidar com bases extensas, desbalanceadas e com variáveis complexas, como ocorre frequentemente em contextos reais de concessão de crédito.

Para tanto, foi utilizada uma base de dados pública da plataforma LendingClub, contendo informações detalhadas sobre solicitações de empréstimos pessoais realizadas entre 2007 e 2018. Após o processo de seleção e pré-processamento das variáveis, três algoritmos supervisionados foram treinados e avaliados: Regressão Logística, Decision Tree e XGBoost. As métricas utilizadas para avaliação dos modelos foram: acurácia, recall, F1-score e AUC-ROC. Elas foram escolhidas por sua capacidade de capturar diferentes aspectos da performance preditiva em bases desbalanceadas.

Os resultados empíricos mostraram diferenças expressivas entre os modelos. A regressão logística, embora amplamente utilizada por sua simplicidade e interpretabilidade, apresentou desempenho limitado em termos de sensibilidade e capacidade discriminativa, indicando suas restrições em cenários mais complexos. O modelo de Decision Tree apresentou a maior acurácia bruta, mas com recall e F1-score significativamente baixos, o que compromete sua utilidade prática no contexto de risco de crédito, em que identificar corretamente os inadimplentes é relevante. Já o XGBoost foi o modelo com melhor desempenho global, apresentando resultados superiores em todas as métricas, com destaque para o AUC-ROC e o recall, demonstrando sua eficiência na detecção de padrões complexos e sua adequação à realidade dos dados financeiros. Esses achados corroboram a literatura, segundo a qual modelos baseados em técnicas de aprendizado de máquina, especialmente aqueles que utilizam métodos de ensemble, como o XGBoost, têm se mostrado mais eficazes do que abordagens tradicionais quando aplicados a problemas de classificação em ambientes com alta dimensionalidade e desbalanceamento de classes (Lai et al., 2022).

Dessa forma, os resultados deste estudo contribuem tanto para a academia quanto para o mercado financeiro, ao demonstrar a viabilidade e o potencial dos modelos de Machine Learning na modernização da análise de crédito. Além de permitir previsões mais acuradas, essas técnicas possibilitam uma alocação mais eficiente de recursos, com potencial para redução da inadimplência e aperfeiçoamento dos processos de concessão de crédito.

Contudo, é importante reconhecer algumas limitações. A primeira diz respeito ao número de modelos avaliados: embora o XGBoost tenha sido testado com sucesso, ainda não foi possível executar outros algoritmos de alto desempenho, como o Random Forest e o



CatBoost, em razão da exigência computacional desses métodos ao lidarem com grandes bases de dados como a utilizada neste trabalho, o que restringe parcialmente a comparação entre técnicas de ensemble. Além disso, todas as análises foram feitas com base em uma única base de dados. A replicação dos resultados com diferentes conjuntos de dados, de outras instituições financeiras e com outros perfis de tomadores de crédito, poderia fortalecer as conclusões apresentadas.

Em síntese, este trabalho mostrou que os algoritmos de aprendizado de máquina são ferramentas promissoras para a avaliação de risco de crédito, apresentando desempenho superior à regressão logística tradicional. Ao avançar nesse campo, espera-se que o uso consciente e técnico desses modelos contribua para a construção de um sistema de crédito mais eficiente, transparente e resiliente.

Referências

- Acuna, E., & Rodriguez, C. (2004, February). The treatment of missing values and its effect on classifier accuracy. In *Classification, Clustering, and Data Mining Applications: Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago, 15–18 July 2004* (pp. 639-647). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Assaf Neto, Alexandre. Mercado financeiro. 13. ed. São Paulo: Atlas, 2016.
- Belle, V., & Papantonis, I. (2021). Principles and practice of explainable machine learning. *Frontiers in Big Data*, 4, 688969.
- Bruce, P., Bruce, A., & Gedeck, P. (2020). *Practical statistics for data scientists: 50+ essential concepts using R and Python*. O'Reilly Media.
- Chang, V., et al. (2024). Credit risk prediction using machine learning and deep learning: A study on credit card customers. *Risks*, 12(11), 174.
- Chang, Y. C., Chang, K. H., & Wu, G. J. (2018). Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Applied Soft Computing*, 73, 914-920.
- Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).
- Coelho, F. F., de Lima Amorim, D. P., & de Camargos, M. A. (2021). Analisando métodos de machine learning e avaliação do risco de crédito. *Revista Gestão & Tecnologia*, 21(1), 89–116.
- Dong, X., et al. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, 14, 241–258.
- Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (3rd ed.). Sebastopol: O'Reilly Media.
- Izbicki, R., & Dos Santos, T. M. (2020). *Aprendizado de máquina: Uma abordagem estatística*. Rafael Izbicki.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer.
- Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9).



- Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2020). *Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies*. MIT Press.
- Kuhn, M., et al. (2013). *Applied predictive modeling*. New York: Springer.
- Lai, S. B. S., et al. (2021). Comparing the performance of AdaBoost, XGBoost, and logistic regression for imbalanced data. *Mathematics and Statistics*, 9(3), 379–385.
- Louzada, F., Ara, A., & Fernandes, G. B. (2016). Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science*, 21(2), 117-134.
- Melsom, B., Vennerød, C. B., & de Lange, P. E. (2022). Explainable artificial intelligence for credit scoring in banking. *Journal of Risk*.
- Naidu, G., Zuva, T., & Sibanda, E. M. (2023). A review of evaluation metrics in machine learning algorithms. In *Computer Science On-line Conference* (pp. 15–25). Cham: Springer International Publishing.
- Narkhede, S. (2018). Understanding AUC-ROC curve. *Towards Data Science*, 26(1), 220–227.
- Nonato, C. T. (2022). *Machine learning aplicado na concessão de crédito: estudo comparativo (Monografia de MBA em Inteligência Artificial e Big Data)*. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos.
- Pereira, P. M. P. (2020). *Análise de risco de crédito usando algoritmos de machine learning (Dissertação de Mestrado)*. Universidade de Lisboa, Portugal.
- Press, S. J., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, 73(364), 699–705.
- Silva, J. S. (2022). *Gerenciamento integrado de riscos: Modelos de predição de risco de crédito em machine learning para a identificação de ativos problemáticos em uma instituição financeira – segmento habitacional PF*. Universidade de Brasília.
- Strauss, E., Júnior, M. V. B., & Ferreira, W. L. L. (2022). A importância de utilizar métricas adequadas de avaliação de performance em modelos preditivos de machine learning. *Projectus*, 7(2), 52–62.
- Supervision, B. (2011). Basel committee on banking supervision. *Principles for Sound Liquidity Risk Management and Supervision (September 2008)*. <https://www.bis.org/publ/bcbs75.htm>.
- Thomas, L., Crook, J., & Edelman, D. (2017). *Credit scoring and its applications*. Society for industrial and Applied Mathematics.
- Venkatesh, B., & Anuradha, J. (2019). A review of feature selection and its methods. *Cybern. Inf. Technol.*, 19(1), 3-26.
- Wang, F., Ross, C. L. (2018). Machine Learning Travel Mode Choices: Comparing the Performance of an Extreme Gradient Boosting Model with a Multinomial Logit Model. *Transportation Research Record*, Vol. 2672(47) 35 – 45.
- Ying, X. (2019). An overview of overfitting and its solutions. In *Journal of Physics: Conference Series* (p. 022022). IOP Publishing.
- Yurdakul, F. (2014). Macroeconomic modelling of credit risk for banks. *Procedia – Social and Behavioral Sciences*, 109(8), 784–793.