FUCAPE PESQUISA E ENSINO S/A – FUCAPE ES

JOÃO LEONOR DO NASCIMENTO SILVA

PREDIÇÃO DE RISCO DE CRÉDITO COM USO DE MACHINE LEARNING PARA A IDENTIFICAÇÃO DE ATIVOS PROBLEMÁTICOS

VITÓRIA
2025

JOÃO LEONOR DO NASCIMENTO SILVA

PREDIÇÃO DE RISCO DE CRÉDITO COM USO DE MACHINE LEARNING PARA A IDENTIFICAÇÃO DE ATIVOS PROBLEMÁTICOS

Dissertação presentada ao Programa de Pós-Graduação em Ciências Contábeis e Administração, da Fucape Pesquisa e Ensino S/A, como requisito parcial para obtenção do título de Mestre em Ciências Contábeis e Administração, – Nível Profissionalizante.

Orientador: Prof. Dr. Roberto Miranda Pimentel Fully

VITÓRIA
2025

JOÃO LEONOR DO NASCIMENTO SILVA

PREDIÇÃO DE RISCO DE CRÉDITO COM USO DE MACHINE LEARNING PARA A IDENTIFICAÇÃO DE ATIVOS PROBLEMÁTICOS

Dissertação apresentada ao Programa de Pós-Graduação em Ciências Contábeis e Administração da Fucape Pesquisa e Ensino S/A, como requisito parcial para obtenção do título de Mestre em Ciências Contábeis e Administração – Nível Profissionalizante.

Aprovada em 03, de setembro de 2025.

BANCA EXAMINADORA

Prof. Dr. Roberto Miranda Pimentel Fully
Fucape Pesquisa e Ensino S/A

Prof.ª Dr. Gabriel Sanfins
Universidade Federal do Rio de Janeiro

Prof. Dr. Octavio Locatelli
Fucape Pesquisa e Ensino S/A

AGRADECIMENTOS

A Deus, minha mais profunda gratidão. Por me conceder a vida, por Sua misericórdia que me sustentou nos momentos difíceis e por Seu amor constante. Agradeço especialmente porque, em muitos momentos, eu mesmo não imaginei que poderia chegar tão longe, mas Ele sempre me ajudou, mesmo quando minhas forças vacilavam. Toda honra e glória sejam dadas a Ele!

À minha esposa, Danielly, minha companheira, que caminhou ao meu lado me encorajando nos momentos mais difíceis. Sua presença foi luz nos dias escuros e força quando eu pensei em desistir. Aos meus filhos, que me enchem de alegria e orgulho, obrigado por compreenderem minhas ausências e por me motivarem a ser melhor a cada dia — este trabalho também é de vocês!

Aos meus pais, pelo exemplo de integridade, dedicação e fé que sempre nortearam minha vida.

Ao meu orientador Dr. Roberto Miranda Pimentel Fully por sua orientação firme, sua generosidade em compartilhar conhecimento e pela confiança no meu potencial desde o início. Agradeço também aos professores, colegas e à equipe do programa de mestrado, que contribuíram para minha formação acadêmica e pessoal com tantas trocas ricas e valiosas.

Aos amigos, que me motivaram a embarcar nessa jornada e aos colegas de trabalho que me apoiaram e orientaram de muitas formas, oferecendo apoio, palavras de encorajamento e atenção nos momentos em que mais precisei — meu muito obrigado!

E a todos que, de alguma forma, fizeram parte dessa caminhada: saibam que levarei comigo a contribuição de cada um. Esta conquista é fruto de um caminho coletivo.

Obrigado, de coração!

“Porque o Senhor dá a sabedoria e da sua boca vem a inteligência e o entendimento”

(Provérbios 2:6)

RESUMO

Este estudo objetivou analisar a eficácia da aplicação de técnicas de *Machine Learning* (ML) para previsão de ativos problemáticos em uma carteira de crédito de capital de giro destinada a pessoas jurídicas, de um banco brasileiro classificado no segmento S1. A pesquisa foi conduzida a partir de uma base histórica composta por 19,2 milhões de registros, dos quais uma amostra de 7 milhões de contratos foi extraída e tratada para modelagem. A variável dependente foi definida conforme os critérios da Resolução CMN nº 4.966/2021, que caracteriza como ativo problemático o contrato com atraso superior a 90 dias ou com fortes indícios de inadimplência. O estudo comparou o desempenho dos algoritmos de *Regressão Logística*, *Random Forest*, *Gradient Boosting* e *XGBoost*, sendo a abordagem de *Random Forest* a que apresentou os melhores resultados com Acurácia 0,9745 , *Recall* de 0,9633 e *F1-Score* de 0,9812. A avaliação da interpretabilidade dos modelos, realizada por meio da metodologia *SHAP*, evidenciou que variáveis como tempo desde a primeira inadimplência, valor da dívida vincenda e prazo remanescente foram os principais determinantes da classificação de risco. Os resultados demonstram que modelos de ML são altamente eficazes na identificação antecipada de contratos com elevado risco de inadimplência, contribuindo para a eficiência na gestão de crédito e a aderência regulatória. A pesquisa reforça a aplicabilidade prática desses modelos em ambientes bancários e destaca a importância da engenharia de atributos e da explicabilidade para garantir a transparência e a confiabilidade das decisões automatizadas. Ao alinhar métodos quantitativos avançados às diretrizes regulatórias, o estudo oferece uma contribuição significativa para a modernização da gestão de risco de crédito no Brasil.

Palavras-chave: risco de crédito; *machine learning*; ativos problemáticos; capital de giro; regulamentação bancária.

ABSTRACT

This study aimed to analyze the effectiveness of applying Machine Learning (ML) techniques to predict problematic assets in a working capital credit portfolio for legal entities at a Brazilian financial institution classified under segment S1. The research used a historical dataset of 19.2 million records, from which a representative sample of 7 million contracts was extracted and processed for modeling. The target variable was defined according to the criteria of Resolution CMN No. 4,966/2021, which considers a financial instrument as a problematic asset when there is a delay of more than 90 days or strong indicators of non-payment. The study compared the performance of the Logistic Regression, Random Forest, Gradient Boosting, and XGBoost algorithms, with the Random Forest approach delivering the best results, achieving an Accuracy of 0,9745%, Recall of 0,9633%, and F1-Score of 0,9812%. Model interpretability was assessed using SHAP (SHapley Additive exPlanations), highlighting variables such as time since first default, outstanding debt value, and remaining contract term as key predictors of credit risk. The results demonstrate that ML models are highly effective in early identification of high-risk contracts, contributing to credit risk management efficiency and regulatory compliance. This research reinforces the practical applicability of ML techniques in banking environments and emphasizes the importance of feature engineering and explainability to ensure transparency and reliability in automated decision-making. By aligning advanced quantitative methods with regulatory frameworks, this study offers a significant contribution to the modernization of credit risk management in Brazil.

Keywords: credit risk; *machine learning*; problematic assets; working capital; banking regulation.

LISTA DE FIGURAS

SUMÁRIO

1 INTRODUÇÃO

O mercado financeiro, como intermediador entre poupadores e investidores (Candelon & Moura, 2024), requer regulação eficaz (Pandey et al., 2022) com isso, Leo et al., (2019) afirma que os modelos de gestão de risco devem ser adaptados a três desafios, expansão regulatória, novas expectativas dos clientes e a emergência de riscos inéditos, essa combinação de fatores requer transformações significativas nos modelos de gestão de risco.

Estudos recentes têm aplicado algoritmos de *Machine Learning* na previsão de inadimplência, como Vieira (2016), que comparou técnicas como *Bagging, Random Forest e AdaBoost* em carteiras habitacionais do Minha Casa Minha Vida, e Silva (2022), que avaliou modelos como Regressão Logística, *Random Forest e Gradient Boosting* para identificar Ativos Problemáticos em contratos habitacionais conforme a Resolução CMN 4.557/2017, utilizando principalmente variáveis do perfil do tomador e características contratuais.

Nesse sentido, conforme estabelece a Resolução nº 4.966 do Conselho Monetário Nacional (CMN, 2021), um instrumento financeiro é caracterizado como ativo problemático, quando apresenta atraso superior a 90 dias no pagamento de principal ou encargos, ou quando existem indicativos de que a obrigação não será integralmente honrada nos termos originalmente pactuados sem a necessidade de recorrer a garantias ou colaterais.

Desse modo, as técnicas de Machine Learning (ML) têm demonstrado crescente potencial na previsão de inadimplência de crédito (Ma et al., 2023). No entanto, sua aplicação em carteiras de capital de giro voltadas ao segmento

varejo ainda não é uma prática amplamente difundida, especialmente em bancos do segmento S1 — instituições com ativos iguais ou superiores a 10% do PIB — que atendem às exigências da Resolução CMN nº 4.966/2021 (BACEN, 2021), voltadas à gestão de ativos problemáticos.

Diante desse cenário, surge a seguinte questão: É possível prever, com antecedência e precisão, a probabilidade de contratos de capital de giro para pessoas jurídicas se tornarem ativos problemáticos, conforme os critérios estabelecidos pela Resolução CMN nº 4.966/2021, por meio de algoritmos de *Machine Learning*?

O objetivo geral deste estudo é desenvolver e validar um modelo de classificação baseado em *Machine Learning* (ML), capaz de prever, com acurácia e antecedência, a probabilidade de um contrato de capital de giro PJ se tornar um ativo problemático, segundo as diretrizes da Resolução CMN nº 4.966/2021. Especificamente, busca-se comparar a performance de diferentes algoritmos — como Regressão Logística, *Random Forest*, *Gradient Boosting* e *XGBoost* — identificar as variáveis mais preditivas e avaliar a eficiência do modelo selecionado para subsidiar a gestão de risco da instituição financeira.

Teoricamente, este trabalho contribui para a literatura sobre gestão de risco de crédito ao aplicar e avaliar metodologias avançadas de análise de dados em um contexto regulatório específico e em uma carteira relevante, abordando a necessidade de métodos mais precisos em cenários de complexidade econômica (Ruan & Jiang, 2024) e atendendo às demandas por inovação impostas pela regulação prudencial (BACEN, 2021). Na prática, a pesquisa justifica-se pelo potencial melhoria na gestão de risco da instituição financeira,

permitindo a identificação precoce de contratos com maior propensão à inadimplência crítica, otimizando a alocação de recursos para ações de mitigação (cobrança, renegociação) e contribuindo para a redução de perdas e a manutenção da estabilidade financeira.

A presente pesquisa foi conduzida utilizando um conjunto de dados composto por 19,2 milhões de registros históricos de contratos de capital de giro de uma instituição financeira brasileira, dos quais foi extraída uma amostra representativa de 7 milhões de contratos após procedimentos rigorosos de limpeza, tratamento e engenharia de atributos. Aplicando técnicas de modelagem preditiva *out-of-time*, os modelos *Random Forest*, *Gradient Boosting, XGBoost* e Regressão Logística foram treinados e avaliados, destacando-se o *Random Forest* como o melhor desempenho com Acurácia 0,9745 , *Recall* de 0,9633 e *F1-Score* de 0,9812.

A análise de importância de variáveis e a interpretação dos resultados com *SHAP* reforçaram a robustez do modelo, evidenciando fatores como tempo desde a primeira inadimplência e valor da dívida vincenda como os principais preditores do risco de crédito. Tais resultados sustentam a proposta da dissertação de promover uma classificação antecipada eficaz de contratos problemáticos, alinhada às exigências da Resolução CMN nº 4.966/2021. Nesse contexto, o modelo desenvolvido configura-se como um sistema dinâmico de monitoramento da carteira, atuando como um *early warning system* capaz de identificar precocemente contratos com alto risco de inadimplência crítica.

2 REFERENCIAL TEÓRICO

2.1. MERCADO FINANCEIRO E O OBJETIVO DE ALOCAÇÃO EFICIENTE DE RECURSOS

O mercado financeiro é essencial na economia moderna, funcionando como um mecanismo complexo e multifacetado para a alocação eficiente de recursos (Mishkin & Eakins, 2018). De acordo com Capeleti et al. (2023), ele atua como o elo entre aqueles que possuem capital excedente, os poupadores, e aqueles que necessitam de recursos para financiar projetos e empreendimentos, os tomadores.

A alocação eficiente de recursos e a mitigação da assimetria informacional configuram-se como funções centrais do sistema bancário, conforme destacado por Agoraki et al. (2022), que ressaltam o papel dos bancos no processamento adequado de riscos e informações, especialmente em razão de sua elevada alavancagem em comparação a outros agentes econômicos. Nessa mesma linha, Burlacu et al. (2023) argumentam que a análise criteriosa do risco de crédito é substancial para orientar os recursos aos setores mais produtivos da economia, sendo essa alocação fortemente influenciada por fatores como o ambiente macroeconômico e a saúde financeira das instituições.

Tang e Sun (2022) acrescentam que a implementação de sistemas robustos de monitoramento pós-concessão de crédito é fundamental para a identificação antecipada de riscos e para a mitigação de perdas.

Complementarmente, Lv et al. (2023) apontam que os bancos enfrentam dificuldades em avaliar com precisão a capacidade de pagamento dos tomadores, sobretudo em cenários marcados por instabilidade política, o que

contribui para o aumento do risco de crédito. Em consonância, Emmanuel et al. (2024) enfatizam que o uso de ferramentas eficazes de avaliação de risco pode melhorar significativamente a lucratividade das instituições financeiras, especialmente em atividades como a análise de solicitações de cartões de crédito e a concessão de empréstimos.

Dando sequência a esse entendimento, Chibane et al. (2024) ressaltam que a atuação dos bancos no fornecimento de crédito não apenas influencia a eficiência da alocação de recursos, mas também promove uma maior integração dos mercados financeiros, especialmente em economias como a da Zona do Euro. Essa integração contribui para ampliar a mobilidade de capitais, otimizando a distribuição de investimentos produtivos e reduzindo a fragmentação financeira entre diferentes regiões.

Em linha com essa perspectiva, Qin et al. (2024) destacam que políticas macroprudenciais, ao controlar a exposição dos bancos ao risco e mitigar o impacto de choques inflacionários, contribuem decisivamente para que a alocação de crédito ocorra de maneira mais criteriosa e sustentável. Segundo os autores, esse processo é fundamental para preservar a estabilidade financeira em economias emergentes, assegurando que os recursos fluam para atividades com maior retorno social e econômico.

Finalmente, Yao e Fan (2025) reforçam a importância de uma análise rigorosa da capacidade de pagamento dos clientes e da avaliação dos riscos envolvidos nas operações de crédito. Considerando que o crédito é um recurso limitado, sua distribuição eficiente é essencial para promover o crescimento econômico. A adequada alocação de crédito pelos bancos comerciais urbanos

pode fomentar diversos setores econômicos locais, impulsionando o desenvolvimento sustentável das regiões. O estudo enfatiza ainda a relevância de políticas creditícias bem estruturadas como mecanismo de fortalecimento do sistema financeiro e de suporte à economia real.

2.2. MERCADO DE CRÉDITO PARA PESSOA JURÍDICA NO BRASIL

A regulação do crédito para pessoas jurídicas (PJ) no Brasil adota uma abordagem multiforme, estruturada por meio de diversas normas emitidas pelo Conselho Monetário Nacional (CMN) e pelo Banco Central do Brasil (BCB), em vez de se concentrar em um único diploma legal. Essa estrutura normativa é complementada por mecanismos que promovem a transparência e a integridade do sistema financeiro, como o Sistema de Informações de Crédito (SCR), regulamentado pela Resolução CMN nº 5.037 (Conselho Monetário Nacional, 2022). O SCR centraliza informações sobre o endividamento dos clientes e, mediante sua autorização, constitui ferramenta essencial para a análise de risco de crédito das pessoas jurídicas.

Sicsú et al., (2020) destaca que o capital de giro é usado para financiar a produção no curto prazo, especialmente via crédito bancário. Já Dash et al., (2023) destacam que os investimentos em capital de giro apresentam maior sensibilidade a restrições financeiras do que os investimentos em ativos fixos, já que empresas financeiramente limitadas tendem a depender mais fortemente de fontes internas de financiamento.

No mesmo sentido, Rosa et al., (2022) em seu artigo evidencia que o capital de giro é substancial para a sustentabilidade das empresas, e sua má

gestão pode levar a problemas financeiros graves, incluindo a falência. A autora argumenta que um nível adequado de capital de giro permite que as empresas honrem obrigações de curto prazo e aproveitem oportunidades de investimento, equilibrando liquidez e rentabilidade.

Já Gallegos Mardones (2022) informa que as empresas latino-americanas enfrentam ambientes macroeconômicos mais voláteis do que as de economias desenvolvidas, o que afeta diretamente suas decisões de investimento em capital de giro. No entanto, o autor explica que o excesso de investimento pode aumentar os custos de financiamento e reduzir a rentabilidade, sugerindo a existência de um nível ótimo de capital de giro.

No Brasil, segundo o BACEN, (2025) o mercado de crédito, principalmente para Pessoa Jurídica (PJ), tem ganhado destaque nos últimos anos, com crescimento de 17,8% em comparação com Janeiro de 2024. E, de acordo com Kedward (2024), a forma como o dinheiro e o crédito são alocados é um fator determinante na definição dos rumos do capitalismo contemporâneo, marcado pela crescente importância das finanças globais.

Também, de acordo com Pereira et al., (2025) o mercado de crédito para Pessoas Jurídicas (PJ) no Brasil apresenta características singulares que o diferenciam dos outros mercados de crédito, pois é um segmento que necessita de atendimento personalizado e de soluções financeiras complexas de acordo com seus desafios.

2.3. REGULAÇÃO DO MERCADO DE CRÉDITO

A Resolução CMN nº 4.966 (Banco Central do Brasil [BACEN], 2021) substituiu a Resolução CMN nº 4.557 (BACEN, 2017), introduzindo um novo paradigma para a gestão de riscos de crédito. Essa regulamentação passou a exigir que instituições financeiras adotem metodologias mais sofisticadas e precisas para avaliar o risco de crédito e constituir provisões adequadas para perdas prováveis, refletindo as melhores práticas internacionais, especialmente no que diz respeito à classificação de créditos, provisionamento, monitoramento de ativos problemáticos e governança (BACEN, 2021).

De acordo com Sousa e Orleans (2022), a Resolução CMN 4.966/2021 representa um marco relevante no arcabouço regulatório do mercado de crédito brasileiro, ao reformular de maneira significativa os critérios para a identificação, mensuração, reconhecimento e baixa de ativos financeiros com risco de crédito elevado. A normativa substitui parcialmente as diretrizes anteriores da Resolução 2.682/1999, modernizando os parâmetros de classificação do risco de crédito e alinhando-os às práticas internacionais de gestão prudencial, com o objetivo de fortalecer a estabilidade e a transparência do sistema financeiro nacional.

O Banco Central do Brasil (Bacen, 2023), por meio do Relatório de Estabilidade Financeira, reforça seu papel na supervisão do Sistema Financeiro Nacional, com o objetivo de assegurar seu funcionamento adequado e mitigar os riscos inerentes às operações de crédito. Nesse contexto, destaca-se a Resolução CMN nº 4.966/2021, que, alinhada às melhores práticas internacionais, evidencia o esforço contínuo do Bacen em aprimorar a

regulamentação do mercado de crédito, com foco na estabilidade financeira e na segurança do sistema. Essa norma estabelece que a avaliação do risco de crédito deve considerar uma gama abrangente de informações, incluindo dados financeiros, informações cadastrais, histórico de crédito, garantias e aspectos setoriais relevantes (Bacen, 2021).

Galvão et al. (2024) destacam que a adoção de regulamentações macroprudenciais exerce papel fundamental na gestão do risco bancário em nível internacional, ao influenciar o comportamento das instituições financeiras e fortalecer a estabilidade do sistema financeiro. Em consonância com essa perspectiva, a Resolução CMN nº 4.966/2021, editada pelo Banco Central do Brasil, promove a adoção de modelos de *stress testing* como ferramenta para avaliar a resiliência das carteiras de crédito diante de cenários adversos, consolidando uma abordagem mais preventiva e proativa na gestão de riscos (Bacen, 2021).

Nesse contexto, de acordo com Pinto et al. (2024) a aplicação de tecnologias emergentes, como *Machine Learning*, permite o desenvolvimento de novos métodos de gerenciamento de riscos. O estudo de García-Céspedes e Moreno (2022) demonstra que técnicas de *Machine Learning* podem ser altamente eficazes na previsão de perdas em carteiras de crédito, replicando com precisão os resultados do modelo generalizado de risco de crédito proposto por Vasicek (1987), amplamente utilizado na indústria financeira.

2.4. MACHINE LEARNING PARA AVALIAÇÃO DO RISCO DE CRÉDITO E MITIGAÇÃO DA ASSIMETRIA INFORMACIONAL

De acordo com Mhlanga (2021), o conceito de aprendizado de máquina foi estabelecido por Arthur Samuel em 1959 como a capacidade das máquinas de aprender com dados sem serem explicitamente programadas para cada tarefa específica. Esse princípio fundamental tem evoluído para se tornar uma das principais ferramentas na avaliação de risco de crédito atualmente.

As técnicas de *Machine Learning* têm demonstrado capacidade notável de reproduzir com elevada precisão os resultados do modelo tradicional de Vasicek (1987), referência por décadas na indústria financeira para avaliação de risco. Estudos recentes apontam que esses novos métodos não apenas igualam, mas frequentemente superam o desempenho dos modelos convencionais.

Bitetto et al. (2022) reforçam essa perspectiva ao destacar que a adoção de modelos baseados em aprendizado de máquina traz benefícios adicionais, como a redução significativa de tempo e custos no processo de avaliação de risco, além de mitigar a assimetria informacional e promover uma alocação mais eficiente dos recursos financeiros.

Já Li et al. (2023) chamam atenção para o papel essencial da análise de grandes volumes de dados no setor bancário contemporâneo, segundo seus estudos, os modelos de *Machine Learning* aplicados a conjuntos de dados abrangentes conseguem prever a probabilidade de inadimplência com uma precisão significativamente maior do que os métodos tradicionais, especialmente quando combinados com técnicas avançadas de processamento em tempo real.

Do mesmo modo, Batchu (2023) complementa essa visão ao destacar que a capacidade de processamento instantâneo de dados permite uma avaliação mais dinâmica e precisa do perfil de risco dos mutuários. De acordo com o autor a agilidade operacional contribui para decisões mais assertivas e para a redução sistemática das perdas por inadimplência, conforme demonstrado em diversos casos práticos.

Em seus estudos, Bitetto et al. (2023) compararam o desempenho de diferentes abordagens para previsão de risco de crédito em PMEs. Os resultados mostraram que modelos como o *Historical Random Forest* superam claramente os métodos tradicionais, alcançando índices de precisão próximos a 90%. Contudo, os autores alertam para desafios importantes, como a dependência de dados históricos consistentes e a necessidade de desenvolver técnicas que garantam a interpretabilidade dos resultados, aspectos particularmente críticos para instituições sob regulação rigorosa.

Da mesma forma, Ruan e Jiang (2024) apresentam evidências de que as instituições financeiras líderes estão cada vez mais adotando técnicas de *machine learning* em seus sistemas de análise de crédito, pois esses modelos, ao incorporarem desde histórico de empréstimos até padrões comportamentais complexos, conseguem monitorar de forma mais eficiente a exposição ao risco creditício.

2.5. HIPOTESES DE PESQUISA

Diante da crescente necessidade de antecipação de riscos em operações de crédito para pessoas jurídicas, especialmente no contexto de capital de giro,

a literatura tem consolidado a importância do uso de indicadores financeiros e comportamentais como instrumentos preditivos. Alagic et al. (2024) destacam que o uso dessas variáveis como preditores de risco de crédito é uma prática amplamente validada e consolidada. Com base nesse referencial, a presente pesquisa se propõe a testar a seguinte hipótese:

H1: Indicadores antecedentes de deterioração financeira ou comportamento de pagamento irregular estão positivamente associados à probabilidade futura de um contrato de capital de giro PJ ser classificado como Ativo Problemático.

No campo da modelagem preditiva, avanços recentes demonstram que métodos baseados em *Machine Learning*, especialmente os não lineares, têm se mostrado mais eficazes para capturar padrões complexos e não triviais em bases de dados com alta assimetria informacional. Bitetto et al. (2023), ao compararem modelos paramétricos e não paramétricos no contexto de risco de crédito de pequenas e médias empresas, constataram que algoritmos como o *Historical Random Forest* superam modelos tradicionais em termos de desempenho. Resultados semelhantes foram identificados por Satish et al. (2024) e Zhu et al. (2024), que confirmam a superioridade dos modelos baseados em árvores em diferentes domínios de aplicação. À luz desses achados, a pesquisa formula a segunda hipótese:

H2: Modelos de *Machine Learning* não lineares, como *Random Forest* e *Gradient Boosting Machines*, apresentam desempenho superior aos modelos lineares, como Regressão Logística, na previsão de Ativos Problemáticos.

Além da escolha do modelo, a preparação das variáveis utilizadas no processo de modelagem — conhecida como engenharia de *features* — tem ganhado destaque por seu impacto significativo na performance dos algoritmos. Emmanuel et al. (2024) argumentam que a integração de métodos de seleção de variáveis com base na teoria da informação e classificadores complexos, como *stacked ensembles*, contribui substancialmente para a eficiência e precisão dos modelos preditivos, ao assegurar que apenas os atributos mais relevantes sejam considerados. Com base nessa abordagem, propõe-se a terceira hipótese:

H3: A aplicação de técnicas de engenharia de atributos (features) contribui significativamente para a melhoria da performance dos modelos de *Machine Learning* na classificação de Ativos Problemáticos.

Com base nas evidências empíricas e teóricas discutidas, as hipóteses formuladas neste estudo buscam testar a capacidade preditiva de variáveis financeiras e comportamentais na identificação precoce de contratos com potencial de inadimplência, bem como avaliar o desempenho comparativo entre diferentes abordagens de modelagem — lineares e não lineares — e os ganhos decorrentes da aplicação de técnicas avançadas de engenharia de *features*. A partir da análise dessas três dimensões, espera-se contribuir para o aprimoramento dos modelos de avaliação de risco de crédito voltados ao segmento de capital de giro para pessoas jurídicas, oferecendo subsídios técnicos para decisões mais assertivas e alinhadas à estabilidade e eficiência do sistema financeiro.

3 METODOLOGIA

O presente estudo adota uma abordagem quantitativa, fundamentada em técnicas de Aprendizagem de Máquina (*Machine Learning* - ML), com o objetivo central de desenvolver e validar um modelo preditivo robusto para a identificação antecipada de Ativos Problemáticos (AP) na carteira de crédito de Capital de Giro destinada a Pessoas Jurídicas (PJ).

Este estudo foi conduzido em uma instituição financeira brasileira do segmento S1. O nome da instituição não pode ser revelado devido a cláusulas de sigilo contratual e LGPD, mas os dados foram anonimizados e utilizados com autorização formal para fins de pesquisa acadêmica.

A pesquisa busca responder à questão de como a aplicação de ML pode efetivamente antecipar a classificação de um contrato como AP, aderindo aos parâmetros da Resolução CMN 4.966/2021, de forma a contribuir para o aprimoramento da gestão de risco de crédito e a otimização da alocação de capital na instituição.

A utilização de ML é justificada pela crescente disponibilidade de grandes volumes de dados transacionais e pela capacidade dessas técnicas em identificar padrões complexos para prever riscos potenciais (Alagic et al., 2024). A concepção metodológica geral inspira-se em trabalhos prévios na área de modelagem de risco de crédito com ML, como o de Shelci (2022), contudo, adapta e direciona as técnicas para as particularidades do segmento de crédito PJ para capital de giro.

3.1. FONTE DE DADOS E AMOSTRAGEM

O estudo foi conduzido em uma instituição financeira brasileira classificada no segmento S1 pelo Banco Central do Brasil, considerando seu ativo total superior a 10% do Produto Interno Bruto nacional. Utilizou-se um conjunto de dados históricos, anonimizado, composto por informações de contratos de capital de giro firmados entre janeiro de 2019 e dezembro de 2024.

A base histórica continha 19,2 milhões de registros e 58 variáveis. As variáveis originais estão descritas no Apêndice, na tabela 7.1 incluindo dados cadastrais, contratuais, financeiros e operacionais. Após o processo de limpeza e tratamento, que incluiu remoção de inconsistências, valores ausentes e variáveis com baixa variância, a amostra final foi reduzida para 7 milhões de contratos. O critério de seleção foi preservar a representatividade estatística e assegurar a validade externa dos modelos.

Este cuidado com a representatividade amostral é fundamental para evitar viés de seleção e assegurar a validade externa dos modelos treinados, conforme recomendado por Chang et al. (2024).

3.2. DEFINIÇÃO DA VARIÁVEL ALVO E ENGENHARIA DE ATRIBUTOS

A variável dependente (target) foi definida como binária, classificando os contratos em Ativo Problemático (1) ou Não Ativo Problemático (0), de acordo com os critérios estabelecidos na Resolução CMN nº 4.966/2021, que considera inadimplência superior a 90 dias como fator determinante. A escolha das variáveis finais levou em consideração: (i) relevância teórica segundo a literatura de risco de crédito (i) relevância teórica, conforme evidenciado por estudos que

destacam variáveis com maior impacto preditivo em risco de crédito, como Bosker et al. (2025) e Gasmi et al. (2025); (ii) significância estatística na fase exploratória; e (iii) ausência de vazamento de dados. Variáveis com alta correlação com o desfecho e disponíveis antes do evento de inadimplência foram priorizadas.

Na etapa de engenharia de atributos, foram criadas variáveis adicionais com o objetivo de capturar aspectos dinâmicos dos contratos, tais como idade do contrato em dias, tempo decorrido desde a primeira inadimplência, prazo remanescente, valor da dívida vincenda, valor base de cálculo, quantidade de renegociações, presença de garantias reais, modalidade de crédito e setor econômico do cliente.

Segundo Antar e Tayachi (2025), a engenharia de atributos adequada é uma etapa crítica para maximizar o poder preditivo em modelos de risco de crédito.

A seleção das variáveis preditoras do modelo foi realizada por meio de um processo multifásico, com o objetivo de maximizar o desempenho preditivo e mitigar riscos de sobreajuste e vazamento de dados. Inicialmente, foi definido um conjunto de variáveis candidatas com base na relevância teórica para a análise de risco de crédito. Em seguida, foram excluídas variáveis com variância nula, por não apresentarem poder discriminatório, e foram criadas variáveis derivadas que capturam aspectos temporais relevantes, como a idade do contrato em dias e o tempo desde a primeira inadimplência.

Também foram removidas variáveis com potencial de vazamento de dados, por representarem informações futuras ou consequências da inadimplência, como provisões contábeis, ratings internos e o número de dias em atraso. Por fim, foram descartadas variáveis categóricas com cardinalidade excessiva, a fim de evitar a explosão da dimensionalidade no pré-processamento. O conjunto final, composto por 13 preditores, representa um equilíbrio entre robustez metodológica e capacidade explicativa, servindo como base para a modelagem.

As variáveis finais selecionadas para a modelagem são apresentadas na Tabela 1, contemplando atributos numéricos e categóricos.

Tabela 1- Variáveis Preditivas *(Features*) utilizadas na Modelagem Final e Justificativas

Variável (Nome Original)	**Tipo**	**Justificativa para Inclusão**	**Relação Esperada com Risco (AP)**
Idade Contrato Dias	Numérica	Representa a maturidade do contrato. Contratos muito novos ou antigos podem ter perfis de risco distintos (ex: *seasoning effect*).	Não linear / a investigar
Prazo Contrato Dias	Numérica	Prazo original total da operação. Prazos mais longos podem expor o contrato a mais ciclos econômicos e aumentar o risco.	Positiva (+)
Prazo Remanescente Dias	Numérica	Tempo restante até o vencimento. Um prazo remanescente curto pode diminuir o risco futuro, enquanto um longo o mantém.	Positiva (+)
Quantidade Renegociacao	Numérica	Número de vezes que o contrato foi renegociado anteriormente. Indica histórico de dificuldade/flexibilização.	Positiva (+)
Tempo Desde Primeira Inadimplencia	Numérica	Dias desde o primeiro atraso (limitado a 90, 0 se nunca atrasou). Principal indicador de risco recente.	Positiva (+)
VR Divida Vincenda	Numérica	Saldo devedor que ainda vai vencer. Representa a exposição futura ao risco de crédito para o banco.	Positiva (+)

Variável (Nome Original)	Tipo	Justificativa para Inclusão	Relação Esperada com Risco (AP)
Valor Base Cálculo	Numérica	Valor inicial ou base da operação. Pode estar correlacionado ao porte da operação e ao risco absoluto.	Positiva (+)
Garantia Real	Categórica (Binária 0/1)	Indica a presença (1) ou ausência (0) de garantia real. A ausência de garantia geralmente aumenta o risco percebido.	Negativa (-) para presença (1)
Modalidade	Categórica	Código específico da modalidade (ex: 215, 216). Diferentes modalidades de Capital de Giro podem ter riscos inerentes distintos.	Variável
Num Modalidade	Categórica	Código numérico da modalidade (ex: 2). Similar à Modalidade. (Verificar se não é redundante).	Variável
Num Sub Modalidade	Categórica	Código da submodalidade (ex: 15, 16). Detalha ainda mais o tipo de operação, podendo diferenciar riscos.	Variável
Setor COSIF	Categórica	Código COSIF do setor econômico do cliente (ex: 450). O risco de crédito pode variar significativamente entre setores.	Variável
Tipo Contrato	Categórica	Código do tipo de contrato, 1=Original, 2=Renegociado). Contratos renegociados geralmente possuem maior risco inerente.	Positiva (+) para renegociado

Fonte: Elaborada pelo autor (2025).
Nota: A coluna "Tipo" indica se a variável foi tratada como numérica ou categórica (via *One-Hot Encoding*) no pré-processamento. A coluna "Relação Esperada" indica o impacto esperado do *aumento* da variável na probabilidade de Ativo Problemático (AP), baseado na teoria de risco de crédito (sinal a ser confirmado pela análise SHAP). Para variáveis categóricas, a relação depende da categoria específica.

3.3. PRÉ-PROCESSAMENTO DOS DADOS

O pré-processamento dos dados envolveu a padronização de formatos, conversão de variáveis de data e numéricas, tratamento de valores ausentes com substituição por mediana para variáveis contínuas e moda para variáveis categóricas, transformação de variáveis categóricas via técnica de *One-Hot Encoding*, e padronização de variáveis numéricas utilizando *StandardScaler*.

Foram eliminadas variáveis constantes ou com alta cardinalidade inadequada, conforme boas práticas sugeridas por Emmanuel et al. (2024). Além

disso, foi adotado rigoroso controle para prevenir vazamento de informações (*data leakage*), entendido como a inclusão de variáveis que revelam informações futuras ou pós-evento no momento da previsão, o que pode levar a um desempenho artificialmente elevado do modelo e comprometer sua validade em aplicações reais (Shi et al., 2022). Para mitigar esse risco, foram excluídas variáveis potencialmente contaminadas por conceitos internos de avaliação de risco ou provisões, em consonância com a abordagem recomendada por Černevičienė e Kabašinskas (2024).

3.4. MODELAGEM PREDITIVA

Os algoritmos selecionados para a modelagem foram: Regressão Logística (baseline linear), *Random Forest, Gradient Boosting* e *XGBoost*. A escolha desses modelos é justificada pela ampla adoção em estudos anteriores na literatura de risco de crédito, como os trabalhos de Addo et al. (2018), Bitetto et al. (2023), Satish et al. (2024), Vieira (2016) e Zhu et al. (2024).

As implementações foram realizadas utilizando as bibliotecas *Scikit-learn* e *XGBoost* em ambiente Google Colab, com definição de hiperparâmetros adequada para equilibrar desempenho e interpretabilidade, seguindo recomendações de Beltman et al., (2025).

3.5. TREINAMENTO, VALIDAÇÃO E TESTE

O procedimento de separação dos dados seguiu a estratégia *out-of-time*, com contratos firmados entre 2019 e 2022 utilizados para treinamento e dados de 2023 e 2024 reservados exclusivamente para teste. Essa estratégia visa

garantir a robustez temporal dos modelos e reduzir o risco de *overfitting*, conforme defendido por Shelci (2022) e Vaca et al. (2024).

Durante o treinamento, aplicou-se amostragem estratificada para manter a proporção entre Ativos Problemáticos e Ativos Não Problemáticos, estratégia que reforça a validade da avaliação e é amplamente recomendada em estudos de modelagem preditiva em finanças (Chang et al., 2024).

3.6. AVALIAÇÃO DE DESEMPENHO

Para cada modelo treinado, foram gerados gráficos individuais de matriz de confusão, curva *Receiver Operating Characteristic* (ROC) e curva *Precision-Recall* (PR), salvos em alta resolução para análise detalhada. As métricas de avaliação selecionadas foram: acurácia (*accuracy*), precisão (*precision*), sensibilidade (*recall*), *F1-Score* e área sob a curva ROC (AUC-ROC), além da área sob a curva *Precision-Recall* (AUC-PR).

A escolha do modelo final priorizou o *F1-Score*, seguida pela análise do *recall*, alinhando-se à orientação de priorizar a redução de falsos negativos em ambientes de risco de crédito, como recomendado por Emmanuel et al. (2024) e Beltman, Machado e Osterrieder (2025).

3.7. ANÁLISE DE INTERPRETABILIDADE

A interpretabilidade do modelo *Random Forest*, selecionado como o de melhor desempenho, foi explorada por meio da metodologia *SHAP* (*SHapley Additive exPlanations*). Esta técnica, baseada na teoria dos jogos cooperativos,

permite quantificar a contribuição individual de cada variável para cada predição, oferecendo transparência em modelos complexos (Vaca et al., 2024).

Foram gerados o gráfico *Beeswarm* e o gráfico de Importância Média Absoluta das variáveis, proporcionando uma visão clara do impacto relativo de cada atributo na classificação de risco. A adoção de SHAP é consistente com as melhores práticas de explicabilidade exigidas para aplicações de *Machine Learning* no setor financeiro, conforme descrito por Cerneviciene e Kabasinskas (2024) e Chang et al. (2024).

4 RESULTADOS

4.1. DESEMPENHO DOS MODELOS

Os modelos foram avaliados mediante validação *out-of-time*, com dados de 2019–2022 para treinamento e 2023–2024 para teste, assegurando robustez temporal (Beltman et al., 2025). Foram comparados quatro algoritmos, Regressão Logística, *Random Forest*, *Gradient Boosting* e *XGBoost*, utilizando métricas como acurácia, precisão, *recall*, *F1-Score* e *AUC-ROC*.

A variável alvo Ativo Problemático, utilizada como indicador de risco contratual, revelou que 58,86% dos contratos da amostra são classificados como problemáticos, enquanto 41,14% não apresentam indícios de deterioração. A média dos valores registrados foi de 6,98, com mediana igual a 1, indicando que metade dos contratos problemáticos possui códigos baixos. O desvio padrão de 63,29 evidencia uma ampla dispersão dos dados, reforçada pela presença de valores elevados.

Tabela 2 – Estatísticas Descritivas da Variável Ativo Problemático

Indicador	**Valor**	**Porcentagem**
Ativos não problemáticos	2.880.104	41,14%
Ativos problemáticos	4.119.896	58,86%
Média dos códigos	6,98	-
Mediana dos códigos	1	-
Desvio padrão	63,29	-
Maior código registrado	10.011	-

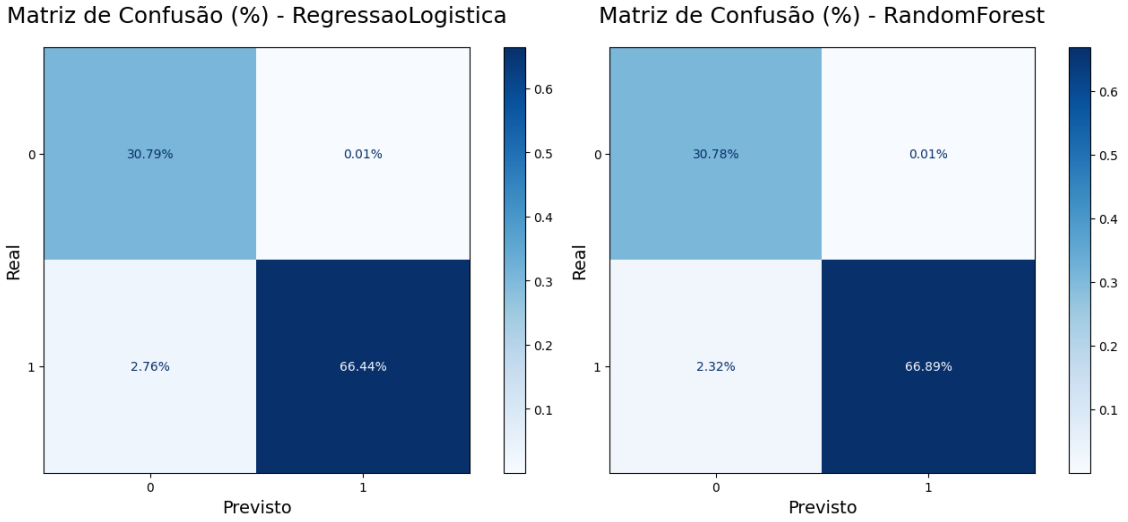
Fonte: Elaborada pelo autor (2025).
Nota: Os dados referem-se à variável Ativo Problemático, utilizada como indicador de risco contratual. A média e o desvio padrão foram calculados sobre todos os registros, incluindo os não problemáticos. As porcentagens indicam a proporção de contratos problemáticos e não problemáticos na amostra total.

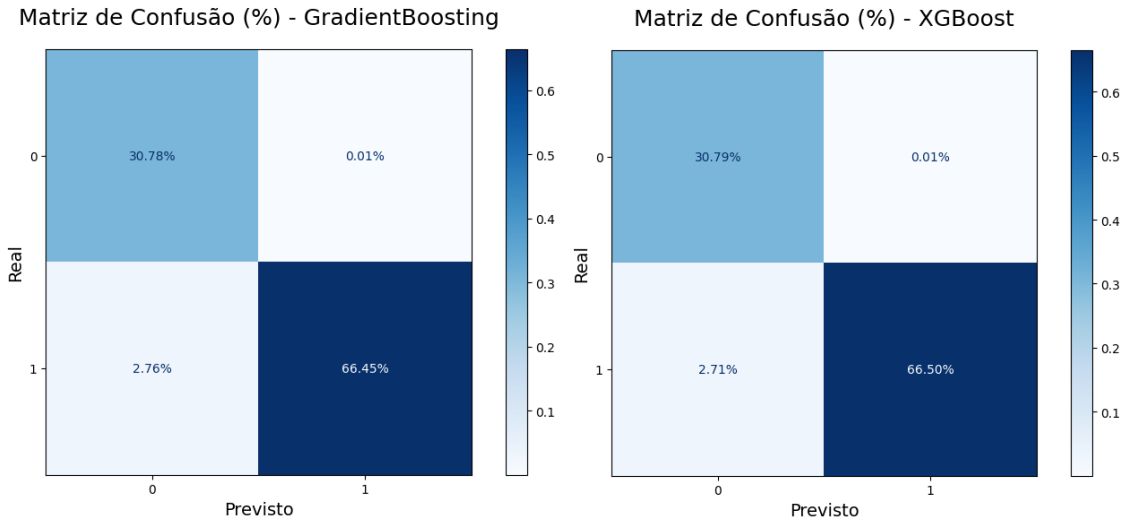
Os resultados obtidos indicaram que todos os modelos apresentaram desempenho elevado, com métricas superiores a 0,95 para todas as avaliações principais. O *Random Forest* destacou-se com Acurácia 0,9745, *Recall* de 0,9633 e *F1-Score* de 0,9812, superando ligeiramente os demais algoritmos em todos os indicadores críticos. Este nível de performance é consistente com evidências recentes da literatura, que apontam a superioridade dos métodos de ensemble em tarefas de previsão de inadimplência em carteiras financeiras complexas (Antar & Tayachi, 2025).

4.2. AVALIAÇÃO GRÁFICA DOS MODELOS

A análise gráfica dos resultados foi conduzida por meio da geração de matrizes de confusão, curvas ROC e curvas *Precision-Recall*. A matriz de confusão do *Random Forest* revelou uma taxa extremamente reduzida de falsos negativos, aspecto fundamental para aplicações de risco de crédito onde a detecção precoce de deterioração é crítica.

Figura 1: Matriz de Confusão da Regressão Logística, *Random Forest*, *Gradient Boosting* e *XGBoost*

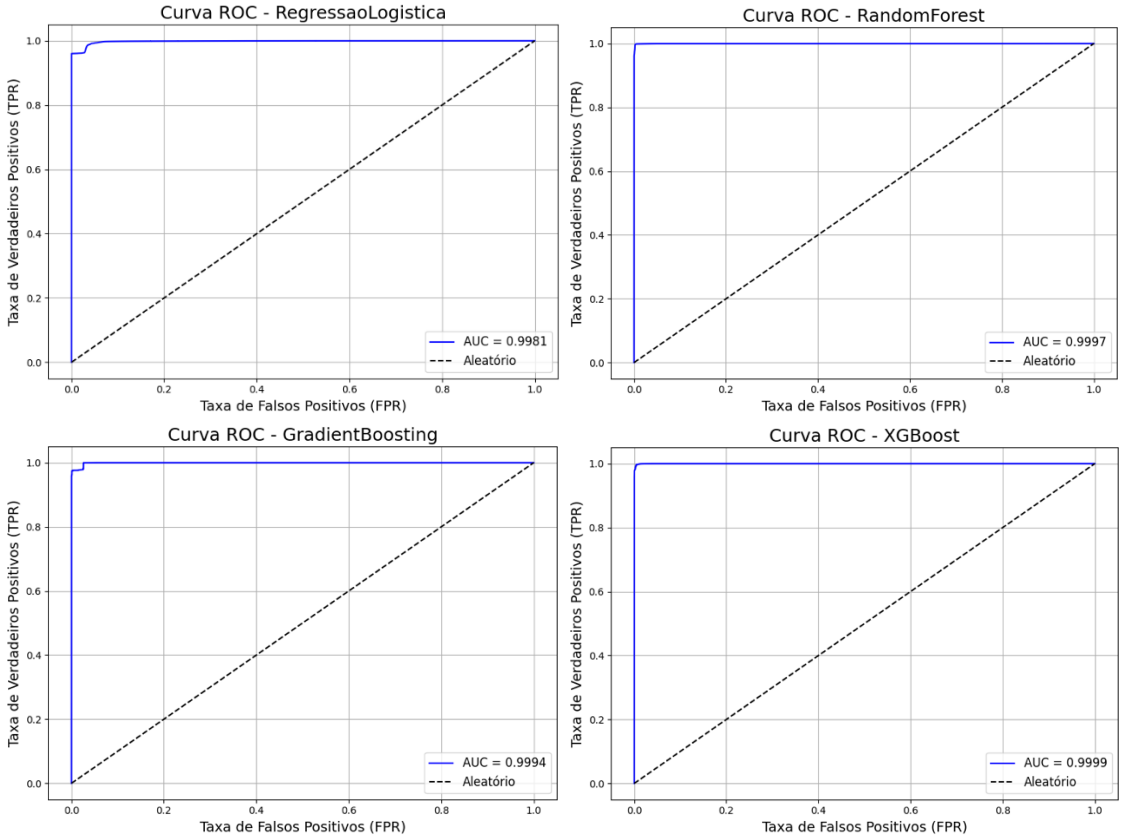




Fonte: Elaborada pelo autor (2025).

As curvas ROC dos modelos evidenciaram áreas sob a curva superiores a 0,99 para todos os algoritmos, com a curva do *Random Forest* praticamente tangenciando o ideal teórico. Este comportamento é coerente com a alta sensibilidade e especificidade observadas.

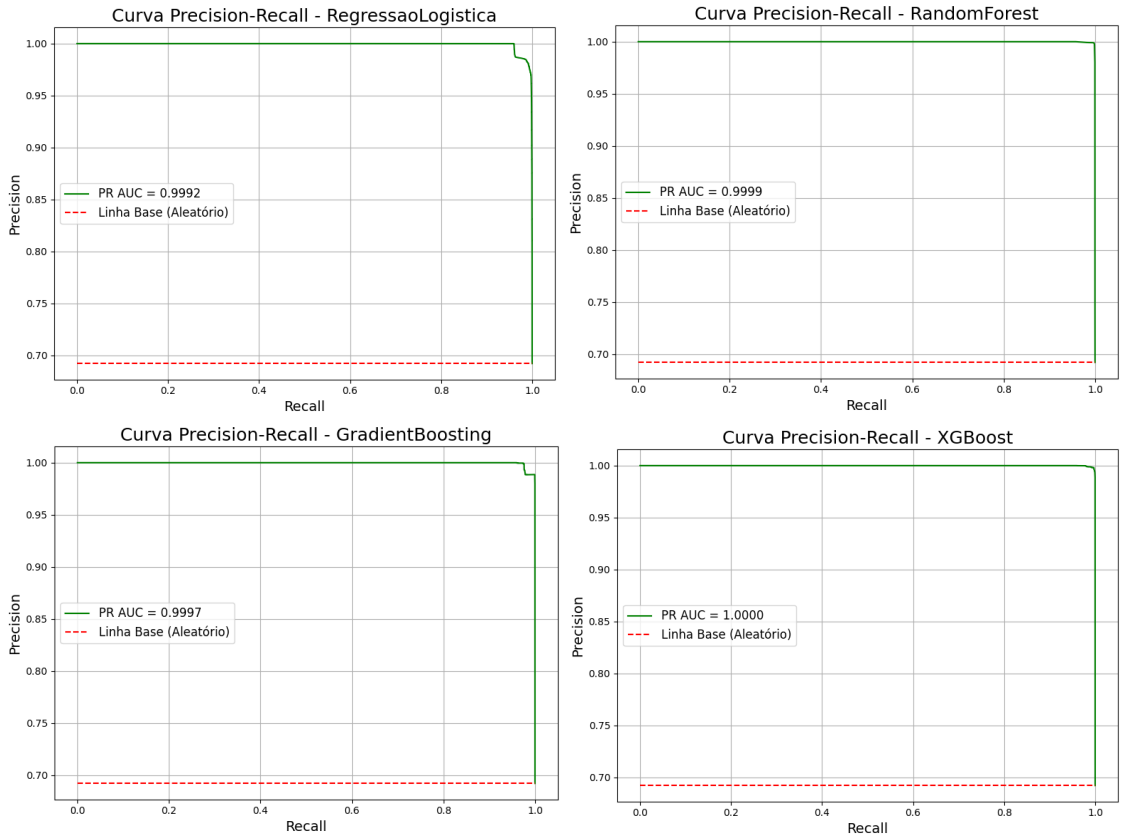
Figura 2: Curva ROC dos Modelos



Fonte: Elaborada pelo autor (2025).

As curvas *Precision-Recall*, mais adequadas para a avaliação em cenários de classes desbalanceadas (Chang et al, 2024), confirmaram a excelente performance dos modelos, em especial do *Random Forest*, que manteve alta precisão mesmo em regimes de elevada sensibilidade.

Figura3 - Curva *Precision-Recall* dos Modelos



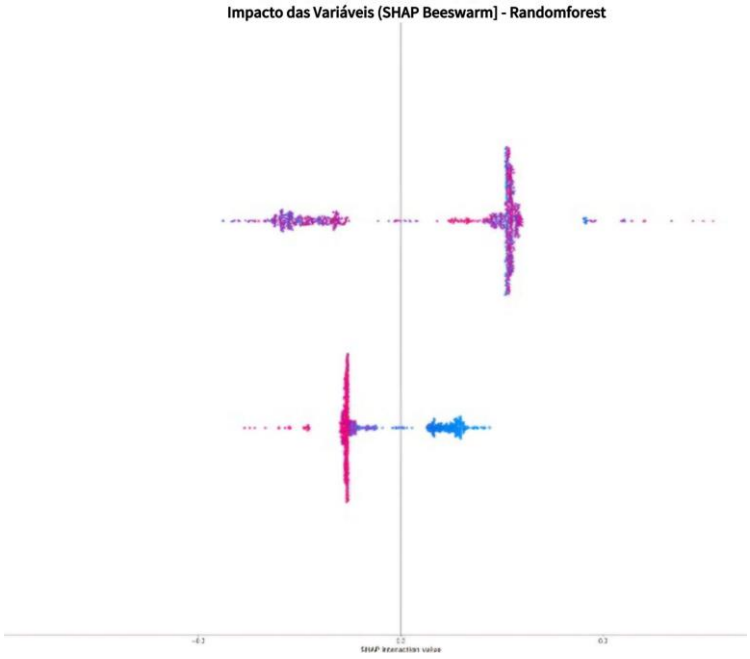
Fonte: Elaborada pelo autor (2025).

4.3. INTERPRETABILIDADE DOS MODELOS

Para garantir a transparência das decisões algorítmicas e atender às exigências da Resolução CMN nº 4.966/2021, foi realizada análise de interpretabilidade utilizando valores *SHAP* (*SHapley Additive exPlanations*). A metodologia *SHAP* permite decompor o impacto de cada variável na decisão final do modelo, proporcionando explicabilidade local e global.

A variável Tempo Desde Primeira Inadimplência destacou-se como principal fator de predição, confirmando a hipótese H1 e corroborando a importância dos indicadores comportamentais recentes na avaliação de risco de crédito (Beltman et al, 2025). O Prazo Remanescente, o Valor da Dívida Vincenda e a Idade do Contrato compuseram o núcleo preditivo secundário, responsáveis conjuntamente por uma fração significativa da capacidade explicativa do modelo.

Figura 4 - Gráfico *SHAP Beeswarm* do *Random Forest* sobre o impacto das variáveis



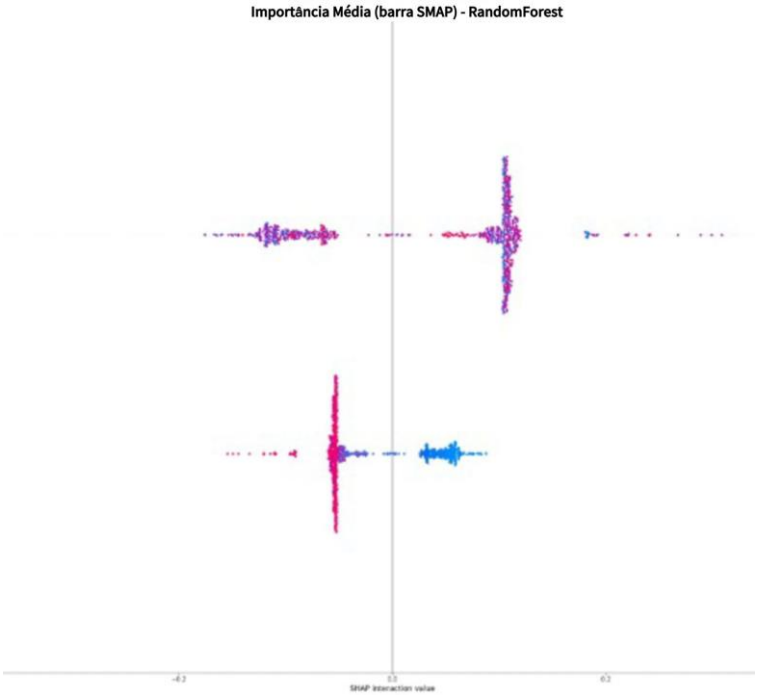
Fonte: Elaborada pelo autor (2025).

A análise visual apresentada na Figura 4 permite compreender de forma granular o impacto individual de cada variável sobre a classificação realizada pelo modelo *Random Forest.* Observa-se que, além de identificar as variáveis com maior influência global, o gráfico *Beeswarm* também revela a direção e a magnitude das contribuições em diferentes observações.

Contratos com valores baixos na variável Tempo Desde Primeira Inadimplência tendem a reduzir a probabilidade de classificação como Ativo Problemático, enquanto valores elevados dessa variável aumentam substancialmente esse risco. Essa abordagem interpretativa reforça a transparência do modelo e valida empiricamente os achados teóricos da literatura, permitindo não apenas a explicação técnica do comportamento algorítmico, mas também subsidiando decisões gerenciais e regulatórias fundamentadas.

Na sequência, a Figura 5 consolida essas informações ao apresentar a importância média das variáveis, auxiliando na priorização dos fatores de risco para ações de mitigação.

Figura 5 - Gráfico de Importância Média dos Valores *SHAP*

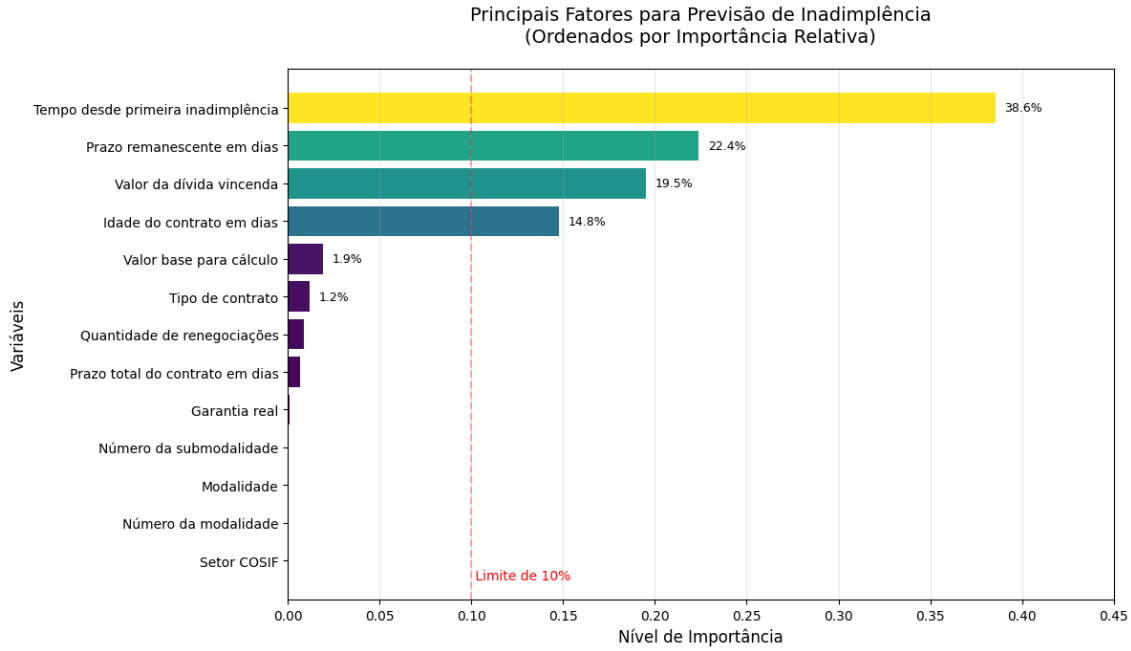


Fonte: Elaborada pelo autor (2025).

Embora a Figura 5 apresente a importância média dos valores *SHAP* com foco na explicabilidade, a Figura 6 complementa a análise ao mostrar a

importância das variáveis segundo o critério interno do *Random Forest*. A convergência entre os dois métodos reforça a robustez do modelo e destaca a consistência dos principais preditores identificados.

Figura 6 - Gráfico de importância das variáveis



Fonte: Elaborada pelo autor (2025).

Variáveis categóricas como tipo de contrato, linha de crédito e existência de garantia real e outras, conforme tabela 1 embora de menor peso relativo, também foram capturadas pelo modelo, validando a hipótese H3 sobre o valor adicional proporcionado pela engenharia de atributos (Černevičienė e Kabašinskas, 2024).

4.4. ESCOLHA DO MODELO FINAL

Com base no conjunto integrado de métricas quantitativas, evidências gráficas e análise de interpretabilidade, o modelo *Random Forest* foi selecionado como o modelo final recomendado para a antecipação de Ativos Problemáticos na carteira de Capital de Giro PJ. A Tabela 3 apresenta o comparativo final de

desempenho dos modelos e a tabela 1 do Apêndice A apresenta as variáveis que compuseram o modelo antes do tratamento de *features*.

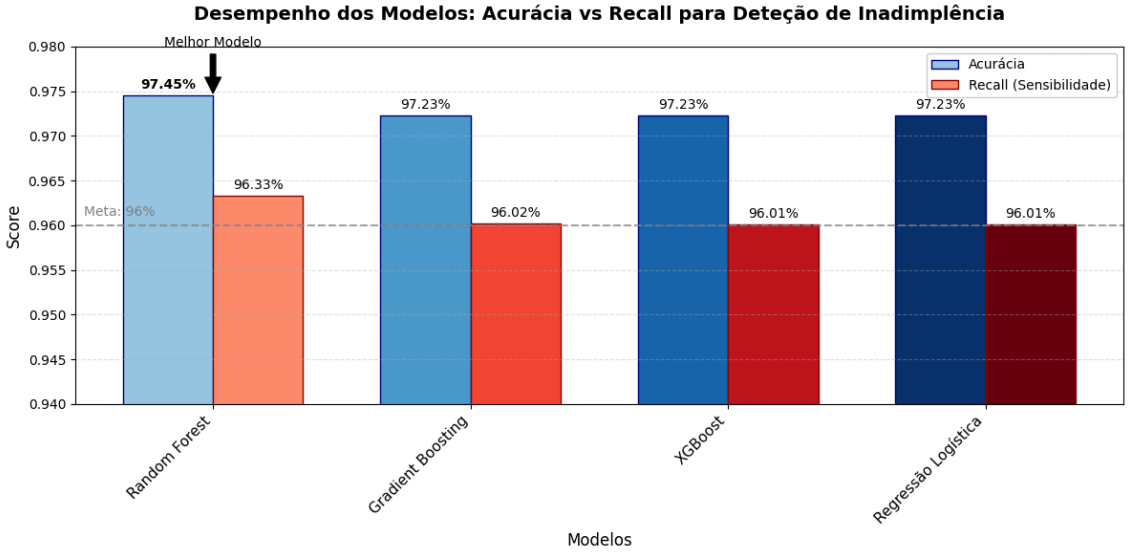
Tabela 3 – Comparativo Final dos Modelos em Porcentagem

Modelo	Acurácia	Precisão	Recall	F1-Score	ROC-AUC
Random Forest	97.45%	99.98%	96.33%	98.12%	99.97%
Gradient Boosting	97.23%	99.99%	96.02%	97.96%	99.94%
XGBoost	97.23%	99.99%	96.01%	97.96%	99.99%
Regressão Logística	97.23%	99.98%	96.01%	97.95%	99.81%

Fonte: Elaboração pelo autor (2025).

A Tabela 3 resume o desempenho dos modelos avaliados com base em métricas tradicionais de classificação. Para facilitar a comparação visual desses resultados, o gráfico a seguir apresenta a performance relativa de cada modelo, permitindo identificar, de forma mais intuitiva, qual algoritmo apresentou os melhores índices em cada métrica.

Figura 7 - Gráfico de desempenho dos modelos quanto a acurácia e Recall (Sensibilidade)



Fonte: Elaborada pelo autor (2025).

A ligeira superioridade do *Random Forest* em termos de Acurácia, *recall*, combinada com sua estabilidade interpretativa e robustez preditiva, justificam sua escolha como solução ótima.

4.5. DISCUSSÃO DOS RESULTADOS

A análise quantitativa apresentada na seção anterior revela um desempenho satisfatório dos modelos de *Machine Learning* aplicados à carteira de Capital de Giro PJ. A capacidade dos algoritmos, em especial do *Random Forest* selecionado, de classificar contratos com propensão a se tornarem Ativos Problemáticos (AP > 90 dias) com métricas de desempenho superiores às alternativas testadas é um dos principais achados desta pesquisa.

A capacidade do *Random Forest* de classificar corretamente contratos que se tornariam Ativos Problemáticos com métricas próximas da perfeição (Acurácia 0,9745 , *Recall* de 0,9633 e *F1-Score* de 0,9812) representa uma contribuição relevante para a literatura e prática de gestão de risco de crédito.

A análise de importância das variáveis, realizada com base na metodologia *SHAP*, oferece percepções valiosas sobre os determinantes do risco na carteira em análise. A predominância da variável Tempo Desde Primeira Inadimplência como o principal fator preditivo corrobora de forma robusta a hipótese H1 formulada no trabalho. Tal resultado está em consonância com a literatura especializada, que enfatiza a relevância de indicadores de alerta precoce no gerenciamento de risco de crédito (Beltman et al., 2025).

Observa-se que a ocorrência recente de inadimplência, ainda que inferior a 90 dias, é um sinal expressivo de deterioração do perfil de risco do cliente. Esse achado implica, do ponto de vista prático, que estratégias de monitoramento focadas em atrasos iniciais (por exemplo, de 1 a 30 dias) podem

ser eficazes para ações preventivas e mitigatórias antes da configuração de um AP pleno, contribuindo para a redução das perdas financeiras.

As variáveis subsequentes em importância, como o Prazo Remanescente, o Valor da Dívida Vincenda e a Idade do Contrato, compõem, em conjunto com o indicador de inadimplência recente, o núcleo preditivo do modelo. Estas variáveis refletem a exposição futura do contrato tanto em termos financeiros quanto temporais, alinhando-se à lógica de avaliação de risco baseada na magnitude e duração da obrigação de pagamento.

Contratos com maior saldo devedor a vencer e maior prazo remanescente naturalmente representam maior exposição e, portanto, maior risco. A variável Idade do Contrato, por sua vez, captura o efeito de seasoning, que sugere que contratos recém-originados ou excessivamente antigos podem apresentar padrões distintos de risco, conforme observado em estudos anteriores (Antar & Tayachi, 2025).

Embora com impacto menor, a utilização de variáveis categóricas transformadas via *One-Hot Encoding* permitiu a identificação de efeitos secundários relevantes, validando parcialmente a hipótese H3 relacionada aos ganhos proporcionados pela engenharia de atributos.

O papel da engenharia de atributos, postulado na Hipótese H3, foi validado quantitativamente por meio de um estudo de ablação. A comparação entre um modelo treinado com e sem a engenharia de *features* Idade do Contrato em Dias e Tempo Desde a Primeira Inadimplência demonstraram um ganho de performance (F1-Score médio de 0.9993 vs. 0.9977) que se provou

estatisticamente significativo (p = 0.031). Este resultado comprova que a transformação de dados brutos em variáveis com significado de negócio foi um pilar fundamental para o sucesso preditivo do modelo, superando o impacto de variáveis categóricas como tipo de contrato ou existência de garantias.

Variáveis como o tipo de contrato, original ou renegociado, o produto de crédito associado e a existência de garantias reais foram reconhecidas como elementos adicionais no processo de classificação, ainda que sua contribuição relativa tenha sido inferior aos fatores quantitativos dominantes.

A baixa influência da variável Quantidade de Renegociações pode ser atribuída à forte capacidade preditiva já capturada pelo Tempo Desde Primeira Inadimplência. Assim, o comportamento recente do cliente se mostrou mais informativo do que o histórico acumulado de renegociações passadas, reforçando a necessidade de abordagem dinâmica na gestão do risco de crédito.

Apesar dos resultados promissores, é fundamental ressaltar que o desempenho do modelo, embora elevado, não alcançou métricas de perfeição, fato que é, inclusive, desejável para evitar indícios de *overfitting*. Conforme argumentado por Chang et al. (2024), métricas extremamente altas em validações podem mascarar problemas de generalização. No presente caso, o uso de validação temporal (*out-of-time*) confere maior robustez às conclusões, mas recomenda-se que o modelo seja continuamente monitorado e testado em novos períodos econômicos para validar sua estabilidade ao longo do tempo.

A base de teste utilizada, composta pelos anos de 2023 e 2024, apresentou distribuição de classes compatível com o esperado para carteiras

corporativas de crédito, sem a concentração excessiva de casos problemáticos que poderia artificialmente inflar a performance dos algoritmos. Isso aumenta a confiança nos resultados obtidos e reforça a aplicabilidade do modelo desenvolvido para o suporte às práticas de gestão de risco conforme estabelecido na Resolução CMN nº 4.966/2021.

Em conclusão, a performance observada indica que a aplicação de algoritmos de *Machine Learning*, em especial modelos baseados em *Random Forest*, é adequada para a antecipação de ativos problemáticos na carteira de Capital de Giro PJ. A identificação de variáveis críticas, o equilíbrio entre sensibilidade e especificidade e a capacidade de explicação dos modelos reforçam a viabilidade de sua utilização prática para a otimização de estratégias de mitigação de risco e gestão ativa da carteira de crédito.

4.6 TESTES DE HIPÓTESES ESTATÍSTICAS APLICADAS

Com o objetivo de aprofundar a avaliação do modelo para além das métricas preditivas usuais, foram conduzidos testes estatísticos que buscavam validar a robustez dos atributos selecionados, bem como comparar o desempenho entre diferentes abordagens algorítmicas e estratégias de engenharia de atributos. Para isso, foram formuladas três hipóteses principais, testadas por meio de métodos estatísticos apropriados.

A primeira hipótese (H1) visou investigar a existência de associação estatisticamente significativa entre a variável derivada Tempo Desde a Primeira Inadimplência e a variável alvo, correspondente ao status de Ativo Problemático. Utilizou-se o coeficiente de correlação bisserial-ponto, apropriado para avaliar

relações entre variáveis contínuas e dicotômicas. O resultado obtido foi um coeficiente de correlação r=0,6782, com valor de p<0,0001, indicando uma associação positiva forte e estatisticamente significativa. Esse achado confirma a relevância dessa variável como importante indicadora de risco de inadimplência no curto prazo.

A segunda hipótese (H2) teve como propósito verificar se o algoritmo *Random Forest* apresentava desempenho superior, em termos de *F1-score*, quando comparado ao modelo de Regressão Logística. Para tal, foi utilizada uma subamostra estratificada composta por 100.000 registros, sobre a qual foi aplicada validação cruzada estratificada em cinco dobras (*folds*). Os resultados dos modelos foram comparados por meio do teste de *Wilcoxon* para amostras pareadas. O teste revelou um valor de p=0,0313, indicando que a superioridade do *Random Forest* em relação à Regressão Logística é estatisticamente significativa no contexto da amostra analisada.

A terceira hipótese (H3) investigou os impactos da inclusão de variáveis derivadas, em particular, aquelas relacionadas ao tempo desde eventos de inadimplência e à idade do contrato sobre a performance do modelo. Foram comparadas duas versões do modelo, uma contendo todas as variáveis, incluindo as derivadas, e outra composta apenas pelas variáveis base. A análise, realizada também com o teste de *Wilcoxon*, indicou um valor de p=0,0078, evidenciando que a inclusão das variáveis derivadas resultou em ganhos estatisticamente significativos de desempenho preditivo.

Os resultados obtidos com esses testes reforçam a robustez estatística do modelo proposto. As evidências empíricas sustentam tanto a validade das

variáveis selecionadas quanto a escolha do algoritmo *Random Forest* como ferramenta preditiva adequada para a tarefa de identificação de ativos problemáticos em ambiente operacional real.

4.7 DISCUSSÃO SOBRE A APLICABILIDADE DO MODELO E A PREVISIBILIDADE DAS VARIÁVEIS

A análise dos resultados obtidos requer uma reflexão acerca da aplicabilidade prática do modelo e de seu desempenho preditivo. Notadamente, a variável Tempo Desde a Primeira Inadimplência foi identificada como o principal fator explicativo, conforme evidenciado pela análise baseada em *SHAP*. Essa constatação conduz a uma indagação metodológica essencial: estaria o modelo utilizando indícios de causalidade reversa ou informações indisponíveis em um ambiente preditivo real, configurando um vazamento de dados (*data leakage*) sutil?

Considera-se que, embora a variável Tempo Desde a Primeira Inadimplência tenha se mostrado altamente preditiva nos testes realizados, sua presença reflete um cenário analítico, e não operacional. Em situações reais de concessão de crédito, essa informação não estaria disponível para todos os contratos, o que sugere a necessidade de adaptar o modelo para ambientes prospectivos com variáveis causalmente consistentes.

A resolução dessa questão está diretamente relacionada ao propósito para o qual o modelo foi concebido: trata-se de um sistema dinâmico de monitoramento da carteira (*early warning system*), em contraste com um modelo de originação de crédito. No contexto de um modelo de concessão, essa variável

seria, por definição, inválida, visto que o contrato ainda não possui histórico de pagamento.

Todavia, o presente estudo assume uma abordagem distinta. O modelo é executado em um momento *t*, com o objetivo de identificar, entre os contratos ativos, aqueles com probabilidade de evoluir para Ativos Problemáticos em um horizonte futuro *t+n* (tipicamente 30 a 60 dias). Nesse contexto de monitoramento, o estado atual do contrato, incluindo a quantidade de dias em atraso, representa uma informação válida, acessível e legítima no instante da previsão.

Consequentemente, o modelo não incorre em vazamento temporal, ao contrário, ele demonstra elevado potencial ao explorar a evidência de risco mais robusta, o padrão recente de inadimplência. A elevada taxa de acerto na predição de que contratos com 80 dias de atraso se tornarão problemáticos não constitui uma tautologia, mas sim a confirmação de que o algoritmo internalizou com precisão a regra fundamental do risco de crédito, isto é, a persistência da inadimplência.

A literatura especializada tem demonstrado, de forma consistente, a relevância de variáveis comportamentais como preditores de risco de crédito. Estudos empíricos em modelagem de risco indicam que variáveis recentes, como inadimplência e exposição financeira corrente, são especialmente eficazes em horizontes de curto prazo. Bitetto et al. (2024), por exemplo, demonstraram que, mesmo com acesso restrito a dados históricos, algoritmos de *Machine Learning* conseguem prever com alta acurácia o risco de crédito de pequenas empresas a partir de variáveis como *Delinquency* e *Outstanding*, que

apresentaram os maiores valores SHAP em seus modelos. Essa evidência é corroborada por Gasmi et al. (2025), que identificaram indicadores comportamentais como determinantes na classificação de clientes, e por Noriega et al. (2023), que destacam em revisão sistemática a recorrência dessas variáveis em modelos preditivos de microfinanças.

Dessa forma, a principal contribuição prática do modelo reside não na predição de eventos raros ou de cauda longa, mas na automatização, priorização e escalabilidade do processo de monitoramento de risco. Em grandes carteiras de crédito, o modelo oferece um mecanismo robusto para direcionar recursos de cobrança e renegociação com proatividade, orientando ações para contratos cuja trajetória de deterioração é clara e mensurável, conforme capturado pela variável Tempo Desde a Primeira Inadimplência.

Essa abordagem está alinhada com os princípios dos sistemas de alerta precoce (*early warning systems*), que visam antecipar sinais de deterioração e permitir respostas ágeis e direcionadas (Akhamere, 2024; Wang, 2024). A literatura especializada também destaca que modelos baseados em *Machine Learning* são especialmente eficazes em ambientes de crédito pulverizado, por sua capacidade de lidar com grandes volumes de dados e gerar insights acionáveis para a gestão de risco (Noriega et al., 2023).

5 CONSIDERAÇÔES FINAIS

O presente trabalho teve como objetivo desenvolver e validar um modelo preditivo, baseado em técnicas de *Machine Learning*, para antecipar a deterioração de ativos de crédito em carteiras de instituições financeiras, em conformidade com a Resolução CMN nº 4.966/2021. A investigação buscou responder à viabilidade de prever, com precisão e robustez, a transição de contratos para a condição de ativos problemáticos, a partir de variáveis históricas e comportamentais do tomador.

Os experimentos conduzidos demonstraram que algoritmos como o *Random Forest*, aliados a um processo criterioso de engenharia de atributos, são capazes de identificar padrões significativos associados ao risco de inadimplência. Destacou-se, entre as variáveis mais relevantes, o tempo desde a primeira inadimplência, cuja importância foi corroborada pela análise SHAP. Longe de representar uma redundância preditiva, essa variável reflete a persistência do comportamento inadimplente, já documentada na literatura e reforçada pelos resultados desta pesquisa.

Diferentemente de modelos de concessão de crédito, o modelo proposto tem foco no monitoramento contínuo de contratos ativos, com previsões realizadas em janelas móveis, 30 a 60 dias à frente. Nesse contexto, todas as variáveis utilizadas são legítimas e disponíveis no momento da previsão. Para assegurar a generalização temporal dos resultados, foi realizada validação *out-of-time*, utilizando dados de períodos posteriores ao treinamento, o que reforça a robustez e aplicabilidade do modelo em cenários prospectivos.

A principal contribuição prática do trabalho está na construção de um sistema de alerta precoce, que oferece suporte à tomada de decisão em processos de cobrança e renegociação. Em carteiras com grande volume, tal mecanismo permite priorizar esforços com base em risco mensurável, promovendo ganhos de eficiência operacional e mitigação proativa de perdas.

No campo acadêmico, o estudo amplia o entendimento sobre o uso de *Machine Learning* no gerenciamento de risco de crédito, demonstrando sua aplicabilidade em ambientes regulatórios e operacionais. Ao incorporar técnicas de explicabilidade como *SHAP*, contribui ainda para a transparência dos modelos, o que é essencial para sua adoção institucional.

Apesar dos resultados promissores, reconhece-se como limitação o uso de dados provenientes de uma única instituição financeira, o que pode restringir a generalização dos achados. Recomenda-se, como continuidade, a aplicação do modelo em bases heterogêneas, bem como o teste de abordagens mais sofisticadas, como redes neurais profundas. Adicionalmente, sugere-se o avanço na integração de técnicas de explicabilidade modelo-agnósticas e o acompanhamento do impacto do modelo em ambientes reais de tomada de decisão.

Conclui-se, portanto, que o uso estruturado de algoritmos de *Machine Learning* no monitoramento de carteiras de crédito representa uma alternativa viável, eficaz e alinhada às exigências prudenciais atuais, contribuindo para uma gestão mais inteligente e preventiva do risco de crédito no sistema financeiro.

REFERENCIAS

Addo, P. M., Guegan, D., & Hassani, B. (2018). Credit risk analysis using machine and deep learning models. *Risks, 6*(2), 38. https://doi.org/10.3390/risks6020038

Agoraki, M.-E. K., Gounopoulos, D., & Kouretas, G. P. (2022). U.S. banks' IPOs and political money contributions. *Journal of Financial Stability, 63*, 101058. https://doi.org/10.1016/j.jfs.2022.101058

Alagic, A., Zivic, N., Kadusic, E., Hamzic, D., Hadzajlic, N., Dizdarevic, M., & Selmanovic, E. (2024). Machine learning for an enhanced credit risk analysis: A comparative study of loan approval prediction models integrating mental health data. *Machine Learning and Knowledge Extraction, 6*(1), 53–77. https://doi.org/10.3390/make6010004

Alvi, J., & Arif, I. (2024). Credit scorecards & forecasting default events: A novel story of non-financial listed companies in Pakistan. *Asia-Pacific Financial Markets.* https://doi.org/10.1007/s10690-024-09494-3

Antar, M., & Tayachi, T. (2025). Partial dependence analysis of financial ratios in predicting company defaults: Random forest vs XGBoost models. *Digital Finance.* https://doi.org/10.1007/s42521-025-00135-6

Akhamere, G. D. (2024). Real-time credit risk monitoring: AI-driven early warning systems for loan portfolio deterioration. *Iconic Research and Engineering Journals* (IRE), 7(9). *(IRE), 7*(9). https://www.irejournals.com/formatedpaper/1710202.pdf

Banco Central do Brasil. (2017). *Resolução nº 4.553, de 30 de janeiro de 2017*. Estabelece a segmentação do conjunto das instituições financeiras e demais instituições autorizadas a funcionar pelo Banco Central do Brasil para fins de aplicação proporcional da regulação prudencial. https://www.bcb.gov.br/estabilidadefinanceira/exibenormativo?tipo=RESOLU%C3%87%C3%83O&numero=4553

Banco Central do Brasil. (2021). *Resolução CMN Nº 4.966, de 25 de novembro de 2021*. Dispõe sobre os conceitos e os critérios contábeis aplicáveis a instrumentos financeiros, bem como para a designação e o reconhecimento das relações de proteção (contabilidade de hedge) pelas instituições financeiras e demais instituições autorizadas a funcionar pelo Banco Central do Brasil. Banco Central do Brasil. (2023). *Relatório de Estabilidade Financeira,* 22(2). BC. https://www.bcb.gov.br/estabilidadefinanceira/relatorioestabilidade

Beltman, J., Machado, M. R., & Osterrieder, J. (2025). Predicting retail customers' distress in the finance industry: An early warning system approach. *Journal of Retailing and Consumer Services, 82*, 104101. https://doi.org/10.1016/j.jretconser.2024.104101

Bertrand Candelon, & Moura, R. (2024). A multicountry model of the term structures of interest rates with a GVAR. *Journal of Financial Econometrics, 22*(5), 1558–1587. https://doi.org/10.1093/jjfinec/nbae008

Bitetto, A., Cerchiello, P., Filomeni, S., Tanda, A., & Tarantino, B. (2023). Machine learning and credit risk: Empirical evidence from small- and mid-sized businesses. *Socio-Economic Planning Sciences, 90*, 101746. https://doi.org/10.1016/j.seps.2023.101746

Bitetto, A., Cerchiello, P., Filomeni, S., Tanda, A., & Tarantino, B. (2024). Can we trust machine learning to predict the credit risk of small businesses? *Review of Quantitative Finance and Accounting*, 63(3), 925–954. https://doi.org/10.1007/s11156-024-01278-0

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

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

Bosker, M., Gürtler, M., & Zöllner, M. (2025). Machine learning-based variable selection for clustered credit risk modeling. *Journal of Business Economics*, 95(2), 123–145. https://doi.org/10.1007/s11573-024-01213-8

Capeleti, P., Garcia, M., & Sanches, F. M. (2022). Countercyclical credit policies and banking concentration: Evidence from Brazil. *Journal of Banking & Finance, 143*, 106589. https://doi.org/10.1016/j.jbankfin.2022.106589

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Chibane, M., Gabriel, A., & Giménez Roche, G. A. (2024). The impact of bank money on stock market integration: Evidence from the Eurozone. *European Journal of Finance, 30*(18), 2137–2156. https://doi.org/10.1080/1351847X.2024.2355104

Černevičienė, J., & Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: A systematic literature review. *Artificial Intelligence Review, 57*(216). https://doi.org/10.1007/s10462-024-10854-8

Dash, S. R., Sethi, M., & Swain, R. K. (2023). Financial condition, working capital policy and profitability: Evidence from Indian companies. *Journal of Indian Business Research, 15*(3), 318–335. https://doi.org/10.1108/JIBR-12-2020-0378

Emmanuel, I., Sun, Y., & Wang, Z. (2024). A machine learning-based credit risk prediction engine system using a stacked classifier and a filter-based feature selection method. *Journal of Big Data, 11*(1), 23. https://doi.org/10.1186/s40537-024-00882-0

Gallegos Mardones, J. (2022). Gestão do capital de giro e desempenho empresarial: Evidências de empresas latino-americanas. *Economic Research-Ekonomska Istraživanja, 35*(1), 3189–3205. https://doi.org/10.1080/1331677X.2021.1986675

García-Céspedes, R., & Moreno, M. (2022). The generalized Vasicek credit risk model: A machine learning approach. *Finance Research Letters, 47*(Part A), 102669. https://doi.org/10.1016/j.frl.2021.102669

Gasmi, F., Neji, M., Mansouri, F., & Soui, M. (2025). Bank credit risk prediction using machine learning model. *Neural Computing and Applications*, 37(4), 5678–5695. https://doi.org/10.1007/s00521-025-11044-5

Greenwood, R., Landier, A., & Thesmar, D. (2015). Vulnerable banks. *Journal of Financial Economics, 115*(3), 471–485. https://doi.org/10.1016/j.jfineco.2014.11.006

Kedward, K., Gabor, D., & Ryan-Collins, J. (2024). Carrots with(out) sticks: Credit policy and the limits of green central banking. *Review of International Political Economy, 31*(5), 1593–1617. https://doi.org/10.1080/09692290.2024.2351838

Leo, M., Sharma, S., & Maddulety, K. (2019). Machine learning in banking risk management: A literature review. *Risks, 7*(1), 29. https://doi.org/10.3390/risks7010029

Li, L., Lin, J., Ouyang, Y., & Luo, X. (2022). Evaluating the impact of big data analytics usage on the decision-making quality of organizations. *Technological Forecasting and Social Change, 175*, 121355. https://doi.org/10.1016/j.techfore.2021.121355

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774. https://doi.org/10.48550/arXiv.1705.07874

Lv, C., Fan, J., & Lee, C.-C. (2023). Can green credit policies improve corporate green production efficiency? *Journal of Cleaner Production, 397*, 136573. https://doi.org/10.1016/j.jclepro.2023.136573

Ma, Y., Zhang, P., Duan, S., & Zhang, T. (2023). Credit default prediction of Chinese real estate listed companies based on explainable machine learning. *Finance Research Letters, 58*(Part A), 104305. https://doi.org/10.1016/j.frl.2023.104305

Martinez, A. L. (2008). Detectando earnings management no Brasil: Estimando os accruals discricionários. *Revista Contabilidade & Finanças, 19*(46), 7–17. https://doi.org/10.1590/S1519-70772008000100002

Mhlanga, D. (2021). Financial inclusion in emerging economies: The application of machine learning and artificial intelligence in credit risk assessment. *International Journal of Financial Studies, 9*(3), 39. https://doi.org/10.3390/ijfs9030039

Moj, M., & Czarnecki, S. (2024). Comparative analysis of selected machine learning techniques for predicting the pull-off strength of the surface layer of eco-friendly concrete. *Advances in Engineering Software, 195*, 103710. https://doi.org/10.1016/j.advengsoft.2024.103710

Noriega, J. P., Rivera, L. A., & Herrera, J. A. (2023). Machine learning for credit risk prediction: A systematic literature review. *Data*, 8(11), 169. https://doi.org/10.3390/data8110169

Pandey, A., Tripathi, A., & Guhathakurta, K. (2022). The impact of banking regulations and accounting standards on estimating discretionary loan loss provisions. *Finance Research Letters, 44*, 102068. https://doi.org/10.1016/j.frl.2021.102068

Pereira, R., Passari, É. dos S., Santos, S., Torresan, D. de C. M., Silva, P. V. da, & Pimentel, R. E. (2025). O impacto das fintechs e do PIX no atendimento às pessoas jurídicas. *Revista Pesquisa Qualitativa, 13*(33), 49–77. https://doi.org/10.33361/RPQ.2025.v.13.n.33.833

Pinto, A. C., de Carvalho, A. X. Y., Tessmann, M. S., & Lima, A. V. (2024). Are machine learning models more effective than logistic regressions in predicting bank credit risk? An assessment of the Brazilian financial markets. *International Journal of Monetary Economics and Finance, 17*(1), 29–48. https://doi.org/10.1504/IJMEF.2023.10058589

Qin, X., Peng, G., & Zhao, M. (2024). Effects of inflation and macroprudential policies on bank risk: Evidence from emerging economies. International Review of Financial Analysis, 92, 103841. https://doi.org/10.1016/j.irfa.2024.103841

Rosenblatt, M., Tejavibulya, L., Jiang, R., Noble, S., & Scheinost, D. (2024). Data leakage inflates prediction performance in connectome-based machine learning models. *Nature Communications, 15*(1), 1829. https://doi.org/10.1038/s41467-024-46150-w

Ruan, J., & Jiang, R. (2024). Does digital inclusive finance affect the credit risk of commercial banks? *Finance Research Letters, 62*(Part A), 105153. https://doi.org/10.1016/j.frl.2024.105153

Satish, N., Anmala, J., Rajitha, K., & Varma, M. R. R. (2024). A stacking ANN ensemble model of ML models for stream water quality prediction of Godavari River Basin, India. *Ecological Informatics, 80*, 102500. https://doi.org/10.1016/j.ecoinf.2024.102500

Shelci, J. [Silva]. (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* [Dissertação de Mestrado, Universidade de Brasília]. Repositório Institucional da UnB. http://repositorio.unb.br/handle/10482/44727

Shakil, M. H., Pollestad, A. J., & Kyaw, K. (2025). Environmental, social and governance controversies and systematic risk: A machine learning approach. *Finance Research Letters, 106894.* https://doi.org/10.1016/j.frl.2025.106894

Shi, S., Tse, R., Luo, W., D'Addona, S., & Pau, G. (2022). Machine learning-driven credit risk: a systemic review. *Neural Computing and Applications*, 34, 14327–14339. https://doi.org/10.1007/s00521-022-07472-2

Sicsú, J., Modenesi, A. de M., & Pimentel, D. (2020). Severe recession with inflation: The case of Brazil. *Journal of Post Keynesian Economics, 43*(3), 318–341. https://doi.org/10.1080/01603477.2020.1835497

Silva, A. C. S. e, Bragança, G. J. de O. e, Braga, H. de O. L., & Sacramone, M. B. (2022). Classificação do risco das operações de crédito: a resolução 2.682/1999 CMN alterada pela resolução 4.966/2021 CMN: Risk classification of credit operations: resolution 2.682/1999 CMN amended by resolution 4.966/2021 CMN. *Brazilian Journal of Development*, *8*(8), 60031–60047. https://doi.org/10.34117/bjdv8n8-334

Tang, L., & Sun, S. (2022). Fiscal incentives, financial support for agriculture, and urban-rural inequality. *International Review of Financial Analysis, 80*, 102057. https://doi.org/10.1016/j.irfa.2022.102057

Vaca, C., Astorgano, M., López-Rivero, A. J., Tejerina, F., & Sahelices, B. (2024). Interpretability of deep learning models in analysis of Spanish financial text. *Neural Computing and Applications, 36*(7), 7509–7527. https://doi.org/10.1007/s00521-024-09474-8

Vieira, J. R. d. C. (2016). *Predição do bom e do mau pagador no programa minha casa, minha vida* [Dissertação de Mestrado, Universidade Federal de Brasilia]. Repositório Institucional da UnB. http://repositorio.unb.br/handle/10482/22581

Yao, J., & Fan, J. (2025). The impact of policy uncertainty and risk taking on the credit resource allocation of urban commercial banks. *International Review of Economics & Finance, 97*, 103766. https://doi.org/10.1016/j.iref.2024.103766

Wang, Z. Q. (2024). Artificial intelligence and machine learning in credit risk assessment: Enhancing accuracy and ensuring fairness. Open Journal of Social Sciences, 12(11), 12–25. https://doi.org/10.4236/jss.2024.1211002

Zhu, F., Wu, X., Lu, Y., & Huang, J. (2024). Strength estimation and feature interaction of carbon nanotubes-modified concrete using artificial intelligence-based boosting ensembles. *Buildings, 14*(1), 134. https://doi.org/10.3390/buildings14010134

APÊNDICE A -TABELA DAS VARIAVEIS TOTAIS

Tabela 1 - Variáveis totais do banco de dados utilizado no modelo de predição de risco de crédito.

Nome da Coluna	Tipo de Dado	Descrição Inferida
posicao_ano_mes	Número inteiro	Ano e mês de referência do registro
Tipo_Pessoa	Número inteiro	Tipo de pessoa (1 = física, 2 = jurídica)
Num_Sistema	Número inteiro	Código do sistema
Num_Unidade_Concessora	Número inteiro	Unidade que concedeu o contrato
Num_Produto	Número inteiro	Código do produto ou linha de crédito
Numero_Contrato	Número inteiro	Número identificador do contrato
Tipo_Contrato	Número inteiro	Tipo do contrato (ex: CDC, capital de giro, etc.)
Situacao_Contrato	Número inteiro	Situação atual do contrato
Prazo_Contrato_Dias	Número inteiro	Prazo total do contrato em dias
Data_Concessao	Texto / Data	Data em que o crédito foi concedido
Prazo_Remanescente_Dias	Número inteiro	Dias restantes do contrato
Prazo_Atraso_Dias	Número inteiro	Dias de atraso na obrigação
Valor_Base_Calculo	Número decimal	Valor base para cálculo de provisões
Valor_Provisionado	Número decimal	Valor provisionado para perdas
Cod_Avaliacao	Número inteiro	Código da avaliação
Conceito_Avaliacao	Texto	Conceito de avaliação do contrato
Conceito_Avaliacao_Cliente	Texto	Avaliação do cliente (rating interno, por exemplo)
Conceito_Contrato	Texto	Conceito atribuído ao contrato
Cod_Avaliacao_Cliente	Número inteiro	Código da avaliação do cliente
Conceito_contaminado	Texto	Conceito considerando efeito de contaminação (por grupo, por exemplo)
Conceito_conglomerado	Texto	Conceito do conglomerado do cliente
Conceito_arrasto	Texto	Conceito considerando arrasto de risco
Conceito_Provisionamento	Texto	Conceito usado para cálculo de provisão

Indicador_Interface_Provisionamento	Texto	Indicador se houve interface com sistema de provisão
Data_Primeira_Prestacao_Nao_Paga	Texto / Data	Data da 1ª prestação inadimplida
Data_Transf_Compensacao	Texto / Data	Data de transferência de compensação
Setor_COSIF	Número inteiro	Setor segundo código COSIF (sistema contábil do Bacen)
Num_Modalidade	Número inteiro	Código da modalidade de crédito
Num_Sub_Modalidade	Número inteiro	Submodalidade
Num_Entidade_Contabil	Número inteiro	Código da entidade contábil
VR_Divida_Vencida	Número decimal	Valor da dívida vencida
VR_Divida_Vincenda	Número decimal	Valor da dívida ainda a vencer
VR_Renda_Vencida	Número decimal	Valor da receita vencida
VR_Renda_Vincenda	Número decimal	Valor da receita a vencer
VR_CR_PRJ_EXR_ANT	Número decimal	Crédito projetado em exercício anterior
VR_CR_PRJ_EXR_ATU	Número decimal	Crédito projetado no exercício atual
VR_CR_PRJ_EXR_A48	Número decimal	Crédito projetado até 48 meses
VR_CRDTO_LBRR_360	Número decimal	Valor de crédito a liberar até 360 dias
VR_CRDTO_LBRR_A360	Número decimal	Valor de crédito a liberar após 360 dias
VR_LME_CR_VNO_360	Número decimal	Limite de crédito vencido até 360 dias
VR_LME_CR_VNO_A360	Número decimal	Limite de crédito vencido após 360 dias
Indicador_Excepcionalizacao	Texto	Flag de exceção no tratamento do contrato
Situacao_Conceito_Avaliacao	Número inteiro	Situação da avaliação do conceito
Indicador_pessoa_Cong	Número inteiro	Indicador se é pessoa pertencente a conglomerado
CPF_CNPJ_Cong	Número inteiro	CPF ou CNPJ do conglomerado
Modalidade	Número inteiro	Modalidade do crédito
Data_Renegociacao	Texto / Data	Data da renegociação
Garantia_Real	Texto	Tipo de garantia real vinculada
Quantidade_Renegociacao	Número inteiro	Número de renegociações
Data_Ultima_Renegociacao	Texto / Data	Data da última renegociação

Conceito_cliente_puro	Texto	Conceito puro do cliente
Indicador_Situacao_Cura	Número inteiro	Indicador de cura (recuperação da inadimplência)
Data_Inicio_Reestruturacao	Texto / Data	Data de início da reestruturação
Data_Ultima_Reestruturacao	Texto / Data	Última data de reestruturação
Data_Cura_Reestruturacao	Texto / Data	Data em que houve cura na reestruturação
Codigo_Ativo_Problematico	Número inteiro	Indicador se o contrato é um ativo problemático (0 = não, 1 = sim)
Conceito_Cliente_SIRIC	Texto	Conceito do cliente segundo o SIRIC
Id_Pessoa_Contrato	Texto	Identificador do cliente no contrato (hash ou código)

Fonte: Banco de dados da instituição (2025).
Nota: Dados tratados para preservar sigilo.

APÊNDICE B - CÓDIGO PYTHON

```
# -*- coding: utf-8 -*-
"""
Notebook Colab - TREINAMENTO de Modelo de Previsão AP (Capital Giro
PJ)
Versão Final com Validação Estatística e Gráficos
"""
#
==========================================================================
===========
# SEÇÃO DE IMPORTAÇÕES E CONFIGURAÇÕES INICIAIS
#
==========================================================================
===========
#Instalar a Biblioteca SHAP
!pip install shap -q

# Bibliotecas padrão
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import joblib
import os
from datetime import datetime
import time
import gc

# Google Drive
from google.colab import drive

# Scikit-learn, XGBoost e Estatística
from sklearn.model_selection import train_test_split,
StratifiedKFold
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.impute import SimpleImputer
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier,
GradientBoostingClassifier
from sklearn.metrics import (
    accuracy_score, precision_score, recall_score, f1_score,
    roc_auc_score, confusion_matrix, ConfusionMatrixDisplay,
    roc_curve, precision_recall_curve, auc
)
from sklearn.feature_selection import VarianceThreshold
from xgboost import XGBClassifier
```

```
from scipy.stats import wilcoxon, pointbiserialr
import shap

# Configurações e Warnings
import warnings
warnings.filterwarnings('ignore', category=FutureWarning)
warnings.filterwarnings('ignore', category=UserWarning)
pd.set_option('display.max_columns', None)
print("Bibliotecas importadas.")

#
==========================================================================
===========
# ETAPAS 2 e 3: MONTAGEM DO DRIVE E CONFIGURAÇÕES
#
==========================================================================
===========
print("\n--- ETAPA 2 e 3: Montagem do Drive e Configurações ---")
DRIVE_MOUNT_POINT = '/content/drive'
drive.mount(DRIVE_MOUNT_POINT, force_remount=True)

# Configurações Essenciais
DATASET_DRIVE_PATH = 'ARQUIVO DO BANCO DE DADOS
MODEL_SAVE_DIR = 'DIRETORIO'
os.makedirs(MODEL_SAVE_DIR, exist_ok=True)
FILE_SEP = ';'; FILE_DECIMAL = '.'; FILE_ENCODING = 'latin1'
SAMPLE_SIZE = 7000000; RANDOM_STATE_SAMPLE = 42
POSITION_DATE_COL = 'posicao_ano_mes'; CONCESSION_DATE_COL =
'Data_Concessao'
FIRST_DEFAULT_DATE_COL = 'Data_Primeira_Prestacao_Nao_Paga';
DAYS_ARREARS_COLUMN = 'Prazo_Atraso_Dias'
AGE_COLUMN_NAME = 'Idade_Contrato_Dias';
TIME_SINCE_FIRST_DEFAULT_COL_NAME =
'Tempo_Desde_Primeira_Inadimplencia'
TARGET_COLUMN_NAME = 'Ativo_Problematico'
FEATURE_CANDIDATES = [
    'Tipo_Pessoa', 'Num_Sistema', 'Num_Unidade_Concessora',
'Num_Produto',
    'Tipo_Contrato', 'Prazo_Contrato_Dias',
'Prazo_Remanescente_Dias',
    'Valor_Base_Calculo', 'Conceito_contaminado',
'Conceito_conglomerado',
    'Conceito_arrasto', 'Setor_COSIF', 'Num_Modalidade',
'Num_Sub_Modalidade',
    'VR_Divida_Vincenda', 'VR_CRDTO_LBRR_360',
'VR_CRDTO_LBRR_A360',
    'VR_LME_CR_VNO_360', 'VR_LME_CR_VNO_A360',
'Indicador_Excepcionalizacao',
    'Indicador_pessoa_Cong', 'Modalidade', 'Garantia_Real',
```

```
    'Quantidade_Renegociacao',
]
print("Configurações definidas.")

#
# ETAPA 4: CARREGAMENTO DOS DADOS
#
print("\n--- Inicio da ETAPA 4 ---")
df_full = pd.read_csv(DATASET_DRIVE_PATH, sep=FILE_SEP,
decimal=FILE_DECIMAL, encoding=FILE_ENCODING, low_memory=False,
dtype=str)
print(f"Dados completos carregados. Shape inicial:
{df_full.shape}")

#
# ETAPA 5: LIMPEZA E CONVERSÃO DE TIPOS
#
print("\n--- Inicio da ETAPA 5 ---")
for col in df_full.select_dtypes(include=['object']).columns:
df_full[col] = df_full[col].str.strip()
date_cols_to_convert = {
    POSITION_DATE_COL: {'format': '%Y%m', 'errors': 'coerce'},
    CONCESSION_DATE_COL: {'format': '%Y-%m-%d', 'errors':
'coerce'},
    FIRST_DEFAULT_DATE_COL: {'format': '%Y-%m-%d', 'errors':
'coerce'},
}
default_bad_date = pd.Timestamp('1901-01-01')
for col, params in date_cols_to_convert.items():
    if col in df_full.columns:
        df_full[col] = pd.to_datetime(df_full[col], **params)
        if col != POSITION_DATE_COL: df_full[col] =
df_full[col].replace({default_bad_date: pd.NaT})
cols_to_int = ['Tipo_Pessoa', 'Num_Sistema',
'Num_Unidade_Concessora', 'Num_Produto', 'Tipo_Contrato',
'Prazo_Contrato_Dias', 'Prazo_Remanescente_Dias',
'Prazo_Atraso_Dias', 'Setor_COSIF', 'Num_Modalidade',
'Num_Sub_Modalidade', 'Indicador_pessoa_Cong', 'Modalidade',
'Quantidade_Renegociacao']
cols_to_float = ['Valor_Base_Calculo', 'VR_Divida_Vincenda',
'VR_CRDTO_LBRR_360', 'VR_CRDTO_LBRR_A360', 'VR_LME_CR_VNO_360',
'VR_LME_CR_VNO_A360']
for col in cols_to_int + cols_to_float:
    if col in df_full.columns:
        df_full[col] = pd.to_numeric(df_full[col], errors='coerce')
```

```
        if col in cols_to_int: df_full[col] =
df_full[col].astype(pd.Int64Dtype())
if 'Garantia_Real' in df_full.columns:
    df_full['Garantia_Real'] = df_full['Garantia_Real'].map({'S':
1, 'N': 0}).fillna(-1).astype(int)

#
=====================================================================
# ETAPAS 6, 7 e 8: PRÉ-PROCESSAMENTO, AMOSTRAGEM E ENGENHARIA
#
=====================================================================
print("\n--- Iniciando Etapa 6 ---")
constant_columns = [col for col in df_full.columns if
df_full[col].nunique(dropna=True) <= 1]
if constant_columns:
    df_full.drop(columns=constant_columns, inplace=True)
    FEATURE_CANDIDATES = [f for f in FEATURE_CANDIDATES if f not in
constant_columns]

print("\n--- Inicio da ETAPA 7 ---")
if len(df_full) > SAMPLE_SIZE: df = df_full.sample(n=SAMPLE_SIZE,
random_state=RANDOM_STATE_SAMPLE)
else: df = df_full.copy()
del df_full; gc.collect()

print("\n--- Inicio da ETAPA 8 ---")
if POSITION_DATE_COL in df.columns and CONCESSION_DATE_COL in
df.columns:
    df[AGE_COLUMN_NAME] = (df[POSITION_DATE_COL] -
df[CONCESSION_DATE_COL]).dt.days
    if AGE_COLUMN_NAME not in FEATURE_CANDIDATES:
FEATURE_CANDIDATES.append(AGE_COLUMN_NAME)
if POSITION_DATE_COL in df.columns and FIRST_DEFAULT_DATE_COL in
df.columns:
    time_diff = (df[POSITION_DATE_COL] -
df[FIRST_DEFAULT_DATE_COL]).dt.days
    df[TIME_SINCE_FIRST_DEFAULT_COL_NAME] =
np.select([time_diff.isnull(), time_diff < 0, time_diff > 90], [0,
0, 91], default=time_diff).astype(int)
    if TIME_SINCE_FIRST_DEFAULT_COL_NAME not in FEATURE_CANDIDATES:
FEATURE_CANDIDATES.append(TIME_SINCE_FIRST_DEFAULT_COL_NAME)
df.dropna(subset=[DAYS_ARREARS_COLUMN], inplace=True)
df[TARGET_COLUMN_NAME] = (df[DAYS_ARREARS_COLUMN] > 90).astype(int)
potential_derived_feats = [AGE_COLUMN_NAME,
TIME_SINCE_FIRST_DEFAULT_COL_NAME]
derived_feats_in_df = [f for f in potential_derived_feats if f in
df.columns]
if derived_feats_in_df: df.dropna(subset=derived_feats_in_df,
inplace=True)
```

```
#
================================================================================
# ETAPA 8.6 (ADICIONADA): TESTE ESTATÍSTICO PARA HIPÓTESE H1
#
================================================================================
print("\n--- ETAPA 8.6: Teste Estatístico para Hipótese H1 ---")
corr, p_value_h1 =
pointbiserialr(df[TIME_SINCE_FIRST_DEFAULT_COL_NAME],
df[TARGET_COLUMN_NAME])
print(f"Correlação ('{TIME_SINCE_FIRST_DEFAULT_COL_NAME}' vs
Target): r={corr:.4f}, p={p_value_h1:.6f}")
if p_value_h1 < 0.05: print("  Conclusão: Associação
estatisticamente significativa. H1 corroborada.")

#
================================================================================
# ETAPAS 9, 10 e 11: SELEÇÃO, PIPELINES E DIVISÃO
#
================================================================================
print("\n--- Iniciando ETAPA 9: Seleção de Features ---")
features_existing_in_sample = [f for f in FEATURE_CANDIDATES if f
in df.columns]
if POSITION_DATE_COL not in features_existing_in_sample:
features_existing_in_sample.append(POSITION_DATE_COL)
cols_to_drop_for_model = [TARGET_COLUMN_NAME, DAYS_ARREARS_COLUMN,
'Situacao_Contrato', 'Codigo_Ativo_Problematico',
'VR_Divida_Vencida', CONCESSION_DATE_COL, FIRST_DEFAULT_DATE_COL,
'Numero_Contrato', 'Id_Pessoa_Contrato', 'CPF_CNPJ_Cong',
'Data_Transf_Compensacao', 'Num_Entidade_Contabil',
'Indicador_Interface_Provisionamento', 'Data_Renegociacao',
'Data_Ultima_Renegociacao', 'Indicador_Situacao_Cura',
'Data_Inicio_Reestruturacao', 'Data_Ultima_Reestruturacao',
'Data_Cura_Reestruturacao', 'Valor_Provisionado',
'Conceito_Provisionamento', 'VR_Renda_Vencida',
'VR_Renda_Vincenda', 'VR_CR_PRJ_EXR_ANT', 'VR_CR_PRJ_EXR_ATU',
'VR_CR_PRJ_EXR_A48', 'Cod_Avaliacao', 'Conceito_Avaliacao',
'Conceito_Avaliacao_Cliente', 'Conceito_Contrato',
'Cod_Avaliacao_Cliente', 'Conceito_contaminado',
'Conceito_conglomerado', 'Conceito_arrasto',
'Situacao_Conceito_Avaliacao', 'Conceito_cliente_puro',
'Conceito_Cliente_SIRIC', 'Tipo_Pessoa', 'Num_Sistema',
'Num_Produto', 'VR_CRDTO_LBRR_360', 'VR_CRDTO_LBRR_A360',
'VR_LME_CR_VNO_360', 'VR_LME_CR_VNO_A360',
'Indicador_Excepcionalizacao', 'Indicador_pessoa_Cong',
'Num_Unidade_Concessora']
cols_to_drop_for_model = list(set(col for col in
cols_to_drop_for_model if col in features_existing_in_sample and
```

```
col not in [POSITION_DATE_COL, TARGET_COLUMN_NAME,
DAYS_ARREARS_COLUMN]))
features = sorted([f for f in features_existing_in_sample if f not
in cols_to_drop_for_model and f not in [POSITION_DATE_COL,
TARGET_COLUMN_NAME, DAYS_ARREARS_COLUMN]])
if not features: raise ValueError("Nenhuma feature final
selecionada.")
X = df[features].copy(); y = df[TARGET_COLUMN_NAME].copy()

print("\n--- Iniciando ETAPA 10: Definição dos Pipelines ---")
numerical_features =
sorted(X.select_dtypes(include=np.number).columns.tolist())
categorical_features =
sorted(X.select_dtypes(exclude=np.number).columns.tolist())
numeric_transformer = Pipeline(steps=[('imputer',
SimpleImputer(strategy='median')), ('scaler', StandardScaler())])
categorical_transformer = Pipeline(steps=[('imputer',
SimpleImputer(strategy='most_frequent')), ('onehot',
OneHotEncoder(handle_unknown='ignore', sparse_output=True))])
preprocessor = ColumnTransformer(transformers=[('num',
numeric_transformer, numerical_features), ('cat',
categorical_transformer, categorical_features)], remainder='drop')

print("\n--- Iniciando ETAPA 11: Divisão dos Dados OUT-OF-TIME ---
")
train_mask = df[POSITION_DATE_COL] < '2023-01-01'; test_mask =
df[POSITION_DATE_COL] >= '2023-01-01'
X_train, y_train = X.loc[train_mask], y.loc[train_mask]
X_test, y_test = X.loc[test_mask], y.loc[test_mask]
del df; gc.collect()

#
# ETAPA 11.5 (ADICIONADA): TESTES ESTATÍSTICOS PARA H2 e H3
#
print("\n--- ETAPA 11.5: Testes Estatísticos na Subamostra ---")
HYPOTHESIS_TEST_SAMPLE_SIZE = 100000
if len(X_train) > HYPOTHESIS_TEST_SAMPLE_SIZE:
    X_hyp_test, _, y_hyp_test, _ = train_test_split(X_train,
y_train, train_size=HYPOTHESIS_TEST_SAMPLE_SIZE, stratify=y_train,
random_state=RANDOM_STATE_SAMPLE)
else:
    X_hyp_test, y_hyp_test = X_train, y_train
n_splits = 5; skf = StratifiedKFold(n_splits=n_splits,
shuffle=True, random_state=RANDOM_STATE_SAMPLE)

print("\n--- Testando H2: RF vs. LR ---")
```

```
pipeline_rf_test = Pipeline(steps=[('preprocessor', preprocessor),
('classifier', RandomForestClassifier(random_state=42,
class_weight='balanced'))])
pipeline_lr_test = Pipeline(steps=[('preprocessor', preprocessor),
('classifier', LogisticRegression(max_iter=1000, random_state=42,
class_weight='balanced'))])
scores_rf, scores_lr = [], []
for i, (train_idx, val_idx) in enumerate(skf.split(X_hyp_test,
y_hyp_test)):
    print(f"  H2 - Fold {i+1}/{n_splits}...")
    pipeline_rf_test.fit(X_hyp_test.iloc[train_idx],
y_hyp_test.iloc[train_idx]);
scores_rf.append(f1_score(y_hyp_test.iloc[val_idx],
pipeline_rf_test.predict(X_hyp_test.iloc[val_idx])))
    pipeline_lr_test.fit(X_hyp_test.iloc[train_idx],
y_hyp_test.iloc[train_idx]);
scores_lr.append(f1_score(y_hyp_test.iloc[val_idx],
pipeline_lr_test.predict(X_hyp_test.iloc[val_idx])))
stat, p_value_h2 = wilcoxon(scores_rf, scores_lr,
alternative='greater'); print(f"  Resultado H2: Média F1
RF={np.mean(scores_rf):.4f}, Média F1 LR={np.mean(scores_lr):.4f},
p-valor={p_value_h2:.4f}")
if p_value_h2 < 0.05: print("  Conclusão H2: Superioridade do RF é
estatisticamente significativa.")
else: print("  Conclusão H2: Superioridade do RF não é
estatisticamente significativa.")

print("\n--- Testando H3: Engenharia de Atributos ---")
features_base = [f for f in features if f not in [AGE_COLUMN_NAME,
TIME_SINCE_FIRST_DEFAULT_COL_NAME]]
X_hyp_test_base = X_hyp_test[features_base]
preprocessor_base = ColumnTransformer(transformers=[('num',
numeric_transformer, [c for c in numerical_features if c in
features_base]), ('cat', categorical_transformer, [c for c in
categorical_features if c in features_base])], remainder='drop')
pipeline_base = Pipeline(steps=[('preprocessor',
preprocessor_base), ('classifier',
RandomForestClassifier(random_state=42, class_weight='balanced'))])
scores_completo, scores_base = [], []
for i, (train_idx, val_idx) in enumerate(skf.split(X_hyp_test,
y_hyp_test)):
    print(f"  H3 - Fold {i+1}/{n_splits}...")
    pipeline_rf_test.fit(X_hyp_test.iloc[train_idx],
y_hyp_test.iloc[train_idx]);
scores_completo.append(f1_score(y_hyp_test.iloc[val_idx],
pipeline_rf_test.predict(X_hyp_test.iloc[val_idx])))
    pipeline_base.fit(X_hyp_test_base.iloc[train_idx],
y_hyp_test.iloc[train_idx]);
```

```
scores_base.append(f1_score(y_hyp_test.iloc[val_idx],
pipeline_base.predict(X_hyp_test_base.iloc[val_idx])))
stat, p_value_h3 = wilcoxon(scores_completo, scores_base,
alternative='greater'); print(f"  Resultado H3: Média F1
Completo={np.mean(scores_completo):.4f}, Média F1
Base={np.mean(scores_base):.4f}, p-valor={p_value_h3:.4f}")
if p_value_h3 < 0.05: print("  Conclusão H3: Engenharia de
atributos melhora o modelo de forma significativa.")
del X_hyp_test, y_hyp_test; gc.collect()


#
======================================================================
# ETAPA 12: TREINAMENTO E AVALIAÇÃO FINAL
#
======================================================================
print("\n--- Iniciando ETAPA 12: Treinamento e Avaliação dos
Modelos ---")


models = {
    'RegressaoLogistica': LogisticRegression(max_iter=1000,
random_state=42, class_weight='balanced'),
    'RandomForest': RandomForestClassifier(n_estimators=100,
random_state=42, class_weight='balanced', n_jobs=-1),
    'GradientBoosting':
GradientBoostingClassifier(n_estimators=100, random_state=42),
    'XGBoost': XGBClassifier(
        n_estimators=100,
        max_depth=6,
        learning_rate=0.1,
        subsample=0.8,
        colsample_bytree=0.8,
        random_state=42,
        use_label_encoder=False,
        eval_metric='logloss',
        scale_pos_weight= (len(y_train) - sum(y_train)) /
sum(y_train) if sum(y_train) > 0 else 1
    )
}

resultados = pd.DataFrame(columns=['Modelo', 'Accuracy',
'Precision', 'Recall', 'F1-Score', 'ROC-AUC'])
best_pipeline = None
best_model_name = None
best_f1_score = -1

for name, model in models.items():
    print(f"\n=============== Processando: {name} ===============")
```

```
    pipeline = Pipeline(steps=[('preprocessor', preprocessor),
('classifier', model)])

    print("  Iniciando treinamento...")
    start_train_time = time.time()
    pipeline.fit(X_train, y_train)
    train_time = time.time() - start_train_time
    print(f"  Treinamento concluído em {train_time:.2f} segundos.")

    print("  Iniciando avaliação no teste...")
    start_eval_time = time.time()
    y_pred = pipeline.predict(X_test)
    y_proba = pipeline.predict_proba(X_test)[:, 1]
    eval_time = time.time() - start_eval_time
    print(f"  Avaliação concluída em {eval_time:.2f} segundos.")

    # Calcular métricas
    acc = accuracy_score(y_test, y_pred)
    prec = precision_score(y_test, y_pred, zero_division=0)
    rec = recall_score(y_test, y_pred, zero_division=0)
    f1 = f1_score(y_test, y_pred, zero_division=0)
    try:
        roc = roc_auc_score(y_test, y_proba)
    except ValueError:
        roc = 0.0
        print("  Aviso: ROC AUC não pôde ser calculado.")

    print("\n  Métricas:")
    print(f"    Accuracy: {acc:.4f}")
    print(f"    Precision: {prec:.4f}")
    print(f"    Recall: {rec:.4f}")
    print(f"    F1-Score: {f1:.4f}")
    print(f"    ROC-AUC: {roc:.4f}")

    new_result = pd.DataFrame([{'Modelo': name, 'Accuracy': acc,
'Precision': prec, 'Recall': rec, 'F1-Score': f1, 'ROC-AUC': roc}])
    resultados = pd.concat([resultados, new_result],
ignore_index=True)

    if f1 > best_f1_score:
        print(f"  Novo melhor modelo encontrado: {name} com F1-
Score: {f1:.4f}")
        best_f1_score = f1
        best_pipeline = pipeline
        best_model_name = name

    # --- Geração de Gráficos ---
    print("  Gerando gráficos...")
```

```
    # <<< GERAÇÃO DA MATRIZ DE CONFUSÃO  >>>
    try:
        # 1. Deixar a função criar a figura e os eixos
        disp = ConfusionMatrixDisplay.from_predictions(
            y_test,
            y_pred,
            cmap='Blues',
            normalize='all',
            values_format='.2%'
        )

        # 2. Acessar e modificar os elementos do gráfico gerado
        disp.figure_.set_figwidth(8) # Aumentar um pouco o tamanho
da figura
        disp.figure_.set_figheight(6)
        disp.ax_.set_title(f"Matriz de Confusão (%) - {name}",
fontsize=18, pad=20) # Adicionar 'pad' para espaçamento
        disp.ax_.set_xlabel('Previsto', fontsize=14)
        disp.ax_.set_ylabel('Real', fontsize=14)

        # 3. Aplicar tight_layout para ajustar tudo automaticamente
        plt.tight_layout()

        # 4. Salvar e fechar a figura
        save_path_cm = os.path.join(MODEL_SAVE_DIR,
f'matriz_confusao_pct_{name}.png')
        plt.savefig(save_path_cm, dpi=300, bbox_inches='tight')
        print(f"    Matriz de confusão salva em: {save_path_cm}")
        plt.show() # Mostra o gráfico no Colab
        plt.close(disp.figure_) # Fecha a figura para liberar
memória

    except Exception as e:
        print(f"    Erro ao gerar/salvar Matriz de Confusão: {e}")


#
================================================================
# ETAPAS 13, 14, 15, 16: FINALIZAÇÃO
#
================================================================
print("\n--- Iniciando ETAPA 13: Finalização ---")

print("\n--- SCRIPT CONCLUÍDO ---")
print("\n---  ETAPA 13 CONCLUIDA ---")
print("\n...")
print("\n...")
print("\n--- Inicio da ETAPA 14 ---")
```

```
# --- Salvar o Melhor Modelo ---

if best_pipeline is not None:
    try:
        save_path = os.path.join(MODEL_SAVE_DIR,
f'pipeline_treinado_{best_model_name}_amostra{SAMPLE_SIZE}.pkl') #
Adiciona tamanho da amostra ao nome
        joblib.dump(best_pipeline, save_path)
        print(f"\nMelhor pipeline ({best_model_name}) salvo em:
{save_path}")
    except Exception as e: print(f"\nErro ao salvar: {e}")
else: print("\nNenhum modelo para salvar.")


print("\n---  ETAPA 14 CONCLUIDA ---")
print("\n...")
print("\n...")
print("\n--- Inicio da ETAPA 15 ---")
# --- ETAPA 15: Análise de Importância das Features (com ajuste de
tipo de modelo) ---

import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
import os # Ensure os is imported

print("\n--- Iniciando ETAPA 15: Análise de Importância das
Features ---")

# Check if a best pipeline was determined
if 'best_pipeline' in locals() and best_pipeline is not None and
'best_model_name' in locals():
    print(f"Analisando importância para o melhor modelo:
{best_model_name}")
    # Check if the model type supports feature_importances_
    model = best_pipeline.named_steps['classifier']
    if hasattr(model, 'feature_importances_'):
        try:
            # Separar pré-processador e modelo
            preprocessor =
best_pipeline.named_steps['preprocessor']

            # Obter nomes das features pós-pré-processamento
            feature_names_transformed = []
            transformers_list = preprocessor.transformers_

            # Loop through transformers to find 'num' and 'cat'
            num_features = []
            cat_features = []
```

```
            ohe_transformer = None

            for name, trans, cols in transformers_list:
                if name == 'num':
                    num_features = cols
                    feature_names_transformed.extend(num_features)
                elif name == 'cat':
                    cat_features = cols
                    # Check if the transformer is a pipeline (like
in the original code)
                    if isinstance(trans, Pipeline):
                         ohe_transformer =
trans.named_steps.get('onehot')
                    else: # Or directly the OneHotEncoder
                         ohe_transformer = trans

            if ohe_transformer and hasattr(ohe_transformer,
'get_feature_names_out') and cat_features:
                onehot_feature_names =
ohe_transformer.get_feature_names_out(cat_features)

feature_names_transformed.extend(onehot_feature_names)
            elif not cat_features:
                 print("  Nenhuma feature categórica encontrada no
pré-processador.")
            else:
                 print("  Aviso: Não foi possível obter nomes das
features do OneHotEncoder.")
                 # Fallback: generate generic names if needed,
although less informative
                 # feature_names_transformed.extend([f"cat_{i}" for
i in range(model.feature_importances_.shape[0] -
len(num_features))])

            # Ensure the number of names matches the number of
importances
            if len(feature_names_transformed) ==
len(model.feature_importances_):
                # Extrair importâncias
                importances = pd.DataFrame({
                    'feature': feature_names_transformed,
                    'Importance': model.feature_importances_
                }).sort_values(by='Importance',
ascending=False).reset_index(drop=True)

                print(f"\nTop 20 Features (Total
{len(importances)}):")
                print(importances.head(20))
```

```
                # Limpar nomes (ajustado para nova estrutura
get_feature_names_out)
                # OHE names might already be clean 'colname_value'
                # importances['feature_clean'] =
importances['feature'].str.replace('num__', '', regex=False) #
Usually not needed now
                # importances['feature_clean'] =
importances['feature_clean'].str.replace('cat__', '', regex=False)
# Usually not needed now
                importances['feature_clean'] =
importances['feature'] # Start with the original name

                # Tradução opcional de nomes (ADAPTE ESTA LISTA
CONFORME SUAS FEATURES FINAIS!)
                feature_translation = {
                    'Tempo_Desde_Primeira_Inadimplencia': 'Tempo
Desde 1ª Inadimplência',
                    'VR_Divida_Vincenda': 'Valor Dívida Vincenda',
                    'Prazo_Remanescente_Dias': 'Prazo Remanescente
(Dias)',
                    'Idade_Contrato_Dias': 'Idade do Contrato
(Dias)',
                    'Valor_Base_Calculo': 'Valor Base de Cálculo',
                    'Prazo_Contrato_Dias': 'Prazo do Contrato
(Dias)',
                    'Quantidade_Renegociacao': 'Qtde.
Renegociações',
                    # Examples for OHE features (adjust based on
your actual output)
                    'Tipo_Contrato_1': 'Tipo Contrato: 1',
                    'Tipo_Contrato_2': 'Tipo Contrato: 2',
                    'Num_Modalidade_215': 'Modalidade: 215',
                    'Garantia_Real_1': 'Garantia Real: Sim (1)',
                    'Garantia_Real_0': 'Garantia Real: Não (0)',
                    'Setor_COSIF_450': 'Setor COSIF: 450',
                }
                # Apply translation where available, keep original
name otherwise
                importances['Feature Traduzida'] =
importances['feature_clean'].map(feature_translation).fillna(import
ances['feature_clean'])

                # Plotar gráfico das Top 20
                plt.figure(figsize=(12, 8))
                data_to_plot = importances.head(20)
                sns.barplot(
                    data=data_to_plot, # Use only top 20
                    x='Importance', y='Feature Traduzida',
palette='Blues_r'
```

```
                )
                plt.title(f'Top 20 Variáveis Mais Importantes -
{best_model_name}', fontsize=18)
                plt.xlabel('Importância (média de redução de
impureza ou ganho)', fontsize=14) # More generic label
                plt.ylabel('Variáveis (Pré-processadas)',
fontsize=14)
                plt.xticks(fontsize=12)
                plt.yticks(fontsize=12)
                plt.grid(True, axis='x')
                plt.tight_layout()
                save_path_imp = os.path.join(MODEL_SAVE_DIR,
f'grafico_importancia_features_{best_model_name}_top20.png')
                plt.savefig(save_path_imp, dpi=300,
bbox_inches='tight')
                print(f"Gráfico de importância salvo em:
{save_path_imp}")
                plt.show() # Display the plot in Colab

            else:
                 print(f"  Erro: Número de nomes de features
({len(feature_names_transformed)}) não corresponde ao número de
importâncias ({len(model.feature_importances_)}).")

        except Exception as e:
            print(f"Erro na análise de importância: {e}")
            import traceback
            traceback.print_exc() # Print detailed traceback

    else:
        print(f"\nModelo {best_model_name} (tipo:
{type(model).__name__}) não tem o atributo 'feature_importances_'.
Pulando análise de importância baseada em atributos.")
else:
    print("\nNenhum 'best_pipeline' encontrado ou definido. Pulando
ETAPA 15.")

print("\n---  ETAPA 15 CONCLUIDA ---")
print("\n...")
print("\n...")
#
# ETAPA 16: Análise de Interpretabilidade com SHAP
#
try:
    import shap
    shap_available = True
except ImportError:
```

```
    print("Biblioteca SHAP não instalada. Pulando ETAPA 16. Instale
com: pip install shap")
    shap_available = False

import gc
import os # Ensure os is imported

print("\n--- Iniciando ETAPA 16: Análise de Interpretabilidade com
SHAP ---")

# Check if SHAP is installed and a best pipeline exists
if shap_available and 'best_pipeline' in locals() and best_pipeline
is not None and 'best_model_name' in locals():
    print(f"Analisando SHAP para o melhor modelo:
{best_model_name}")
    preprocessor = best_pipeline.named_steps['preprocessor']
    model = best_pipeline.named_steps['classifier']

    is_tree_based = best_model_name in ['RandomForest',
'GradientBoosting', 'XGBoost']

    if is_tree_based:
        try:
            SHAP_SAMPLE_SIZE = min(1000, X_test.shape[0])
            if SHAP_SAMPLE_SIZE <= 0:
                print("  Aviso: Conjunto de teste está vazio.
Pulando análise SHAP.")
            else:
                np.random.seed(RANDOM_STATE_SAMPLE)
                X_shap_sample = X_test.sample(SHAP_SAMPLE_SIZE,
random_state=RANDOM_STATE_SAMPLE)

                print(f"  Aplicando pré-processador à amostra SHAP
de {SHAP_SAMPLE_SIZE} instâncias...")
                X_shap_transformed =
preprocessor.transform(X_shap_sample)

                # Obter nomes das features pós-transformação
                feature_names_transformed =
preprocessor.get_feature_names_out()

                # Criar DataFrame com os dados transformados
(robusto para plotagem)
                if hasattr(X_shap_transformed, "toarray"):
                    X_shap_transformed_df =
pd.DataFrame(X_shap_transformed.toarray(),
columns=feature_names_transformed, index=X_shap_sample.index)
                else:
```

```
                    X_shap_transformed_df =
pd.DataFrame(X_shap_transformed, columns=feature_names_transformed,
index=X_shap_sample.index)

                print("  Criando SHAP TreeExplainer e calculando
valores...")
                explainer = shap.TreeExplainer(model)
                shap_values =
explainer.shap_values(X_shap_transformed)

                # Tratar saída do shap_values para classificação
binária
                if isinstance(shap_values, list) and
len(shap_values) == 2:
                    shap_values_class1 = shap_values[1]
                else:
                    shap_values_class1 = shap_values

                # --- Geração de Gráfico SHAP (Beeswarm) ---
                print("\n  Gerando Gráfico SHAP (Beeswarm)...")
                # <<< MUDANÇA: Ajuste de layout >>>
                plt.figure() # Criar uma nova figura para ter
controle
                shap.summary_plot(shap_values_class1,
X_shap_transformed_df, plot_type="dot", max_display=20, show=False)
                plt.title(f'Impacto das Variáveis (SHAP Beeswarm)
- {best_model_name}', fontsize=16, pad=20)
                plt.tight_layout() # Ajustar layout ANTES de
salvar
                save_path_shap_bee = os.path.join(MODEL_SAVE_DIR,
f'grafico_shap_summary_beeswarm_{best_model_name}.png')
                plt.savefig(save_path_shap_bee, dpi=300,
bbox_inches='tight')
                print(f"    Gráfico SHAP Beeswarm salvo em:
{save_path_shap_bee}")
                plt.show()
                plt.close() # Fechar a figura

                # --- Geração de Gráfico SHAP (Barra) ---
                print("\n  Gerando Gráfico SHAP (Barra)...")
                # <<< MUDANÇA: Ajuste de layout >>>
                plt.figure() # Criar uma nova figura para ter
controle
                shap.summary_plot(shap_values_class1,
X_shap_transformed_df, plot_type="bar", max_display=20, show=False)
                plt.title(f'Importância Média (SHAP Bar) -
{best_model_name}', fontsize=16, pad=20)
                plt.tight_layout() # Ajustar layout ANTES de
salvar
```

```
                save_path_shap_bar = os.path.join(MODEL_SAVE_DIR,
f'grafico_shap_summary_barra_{best_model_name}.png')
                plt.savefig(save_path_shap_bar, dpi=300,
bbox_inches='tight')
                print(f"    Gráfico SHAP Bar salvo em:
{save_path_shap_bar}")
                plt.show()
                plt.close() # Fechar a figura

        except Exception as e:
            print(f"Erro geral na análise SHAP: {e}")
            import traceback
            traceback.print_exc()

    else:
        print(f"\nModelo {best_model_name} não é baseado em árvore
e não é suportado pelo TreeExplainer.")

else:
    if not SHAP_AVAILABLE: print("SHAP não está instalado.")
    else: print("\nNenhum 'best_pipeline' encontrado ou definido.
Pulando ETAPA 16.")

gc.collect()
print("\n--- ETAPA 16 CONCLUÍDA ---")
```