Machine Learning no risco de crédito: Capital de Giro e PRONAMPE

ALESSANDRO MOREIRA DOS SANTOS
Fucape Pesquisa e Ensino S/A

JOÃO LEONOR DO NASCIMENTO SILVA
Fucape Pesquisa e Ensino S/A

ROBERTO MIRANDA PIMENTEL FULLY
Fucape Pesquisa e Ensino S/A

OCTAVIO LOCATELLI
Fucape Pesquisa e Ensino S/A

GABRIEL RODRIGUES BATISTA SANFINS
Universidade Federal Fluminense

Resumo

Este estudo investiga a aplicação de modelos de Machine Learning à predição do risco de crédito corporativo em carteiras com características distintas, Capital de Giro e PRONAMPE. Utilizando dados históricos de uma instituição financeira brasileira classificada no segmento S1, analisam-se duas carteiras padronizadas em 52 variáveis, com definição de ativo problemático alinhada à Resolução CMN nº 4.966/2021 e validação temporal out-of-time. O desenho empírico envolve três experimentos: (i) Baseline, com treino e teste na mesma carteira; (ii) generalização cruzada entre carteiras; e (iii) modelo combinado (pooled) com variável indicadora de origem. O desempenho é avaliado por métricas adequadas a bases desbalanceadas e explicabilidade via valores SHAP. Os resultados indicam robustez intracarteiras, deterioração sob generalização cruzada e diferenças nos determinantes do risco. O modelo combinado apresenta ganhos de eficiência alocativa na priorização do risco, contribuindo para a literatura de generalização de modelos e para a gestão prudencial do risco de crédito.

Palavras-chave: Risco de crédito; Machine Learning; generalização de modelos; crédito corporativo; PRONAMPE.

1. Introdução

O crédito desempenha papel central no funcionamento do sistema financeiro e no crescimento econômico ao viabilizar a intermediação eficiente de recursos entre agentes superavitários e deficitários. Contudo, a expansão do crédito está intrinsecamente associada à adequada mensuração e gestão do risco de inadimplência, elemento fundamental para a estabilidade prudencial das instituições financeiras e para a sustentabilidade do sistema como um todo (Louzis et al., 2012; Ustarz & Fanta, 2021). Nesse contexto, a estimação acurada da probabilidade de *default* constitui um dos pilares da gestão moderna de risco, especialmente após o fortalecimento das abordagens prospectivas de perda esperada introduzidas pelo IFRS 9 e incorporadas à regulação prudencial brasileira.

Avanços recentes em técnicas de *Machine Learning* (ML) ampliaram significativamente as possibilidades de modelagem do risco de crédito, oferecendo ganhos relevantes de desempenho em relação aos métodos estatísticos tradicionais, como a Regressão Logística (Lessmann et al., 2015). Algoritmos não lineares, como Random Forest, Gradient Boosting e XGBoost, têm demonstrado elevada capacidade de capturar padrões complexos e interações entre variáveis financeiras e comportamentais, particularmente em bases de dados amplas e desbalanceadas (Zou & Gao, 2022). Apesar desses avanços, um desafio permanece central tanto na literatura quanto na prática bancária: a capacidade de generalização dos modelos de risco quando aplicados a contextos ou carteiras estruturalmente distintas.

A literatura de *domain adaptation* e *dataset shift* demonstra que mudanças nas distribuições das variáveis explicativas, nos mecanismos geradores do risco ou na composição amostral podem comprometer significativamente o desempenho de modelos preditivos fora do domínio em que foram treinados (Moreno-Torres et al., 2012; Webb et al., 2016). Em ambientes financeiros, esse fenômeno é particularmente relevante, dado que carteiras de crédito diferem substancialmente em termos de perfil dos tomadores, características contratuais, garantias e objetivos econômicos. Modelos desenvolvidos para um determinado segmento podem, portanto, apresentar deterioração significativa de desempenho quando aplicados a outro, mesmo quando utilizam o mesmo conjunto de variáveis (Markov et al., 2022).

No contexto brasileiro, essa questão assume relevância adicional diante da coexistência de linhas tradicionais de crédito corporativo, como o Capital de Giro, e programas de fomento com desenho institucional específico, como o Programa Nacional de Apoio às Microempresas e Empresas de Pequeno Porte (PRONAMPE). Instituído como resposta contracíclica à crise provocada pela pandemia da Covid-19, o PRONAMPE apresenta estrutura diferenciada de

garantias, critérios de elegibilidade e público-alvo, o que sugere padrões de risco potencialmente distintos daqueles observados nas operações convencionais de capital de giro (Araujo et al., 2021; Banco Central do Brasil, 2021). Apesar de sua importância econômica, ainda são escassas as evidências empíricas que avaliam, de forma sistemática, como modelos preditivos de risco se comportam quando transferidos entre essas carteiras.

Diante desse cenário, emerge o seguinte problema de pesquisa: em que medida modelos de *Machine Learning* desenvolvidos para a predição do risco de crédito mantêm sua robustez, estabilidade e capacidade discriminatória quando aplicados a carteiras corporativas com características estruturais distintas, especificamente Capital de Giro e PRONAMPE? Adicionalmente, questiona-se se a integração dessas carteiras em um modelo combinado pode gerar ganhos de eficiência alocativa, preservando a estabilidade e a interpretabilidade exigidas em ambientes prudencialmente regulados.

O objetivo geral deste estudo é avaliar a robustez preditiva, a capacidade de generalização e a eficiência alocativa de modelos de *Machine Learning* aplicados ao risco de crédito corporativo em carteiras distintas. Para tanto, são conduzidos três experimentos complementares: (i) um experimento Baseline, no qual os modelos são treinados e testados dentro da mesma carteira; (ii) um experimento de Generalização Cruzada, que avalia a transferência direta dos modelos entre carteiras; e (iii) um Modelo Combinado (Pooled), que integra as duas bases em um único arcabouço preditivo com variável indicadora de origem.

A pesquisa contribui para a literatura em três dimensões principais. Primeiro, fornece evidências empíricas sobre a robustez preditiva intracarteiras de modelos de *Machine Learning* aplicados ao crédito corporativo, avaliando sua estabilidade sob validações temporais rigorosas. Segundo, investiga as diferenças estruturais de risco entre Capital de Giro e PRONAMPE, por meio da análise comparativa da importância das variáveis explicativas com técnicas de explicabilidade baseadas em valores SHAP. Terceiro, avalia se a modelagem integrada das carteiras promove ganhos de eficiência alocativa, com implicações diretas para a priorização do risco e a governança prudencial de modelos.

Ao explorar empiricamente os limites da generalização entre carteiras heterogêneas e o papel complementar de modelos combinados, o estudo dialoga com a literatura de *domain adaptation* em finanças e oferece subsídios relevantes para a gestão do risco de crédito em instituições financeiras atuantes em ambientes regulados.

2. Referencial teórico e desenvolvimento das hipóteses

A modelagem do risco de crédito é tradicionalmente orientada à estimação da probabilidade de inadimplência, apoiando decisões de precificação, provisões e alocação de capital. No contexto bancário, além do desempenho estatístico, modelos de risco são avaliados quanto à estabilidade temporal, rastreabilidade e capacidade de justificativa, em linha com práticas de governança e supervisão prudencial. Nesse cenário, técnicas de *Machine Learning* (ML) ampliam a fronteira de desempenho ao capturar não linearidades e interações complexas, mas introduzem desafios de generalização entre domínios e de explicabilidade, sobretudo quando aplicadas a carteiras com estruturas de risco distintas (Lessmann et al., 2015; Moreno-Torres et al., 2012; Bussmann et al., 2021).

2.1 Modelagem preditiva do risco de crédito com *Machine Learning*

A Regressão Logística permanece como referência na prática de risco de crédito por sua transparência e compatibilidade com processos de validação e auditoria. Todavia, sua estrutura linear impõe limitações para capturar relações não lineares e efeitos de interação sem engenharia adicional de variáveis, o que tende a reduzir o poder discriminatório em ambientes complexos e heterogêneos (Dong et al., 2012; Vasconcellos de Paula et al., 2019). Em contraste, abordagens baseadas em árvores e *ensembles* — como Random Forest e métodos de *boosting* — favorecem a captura de padrões não lineares e a exploração de estruturas latentes em grandes bases, frequentemente reportando desempenho superior em aplicações de *credit scoring* (Lessmann et al., 2015; Zhu et al., 2019).

Entre os métodos de *boosting*, o Gradient Boosting Machine (GBM) e o Extreme Gradient Boosting (XGBoost) se destacam por seu poder discriminatório e pela capacidade de lidar com desbalanceamento e complexidade preditiva, desde que calibrados para mitigar sobreajuste (Zou & Gao, 2022; Li et al., 2022). Em aplicações de risco, esse *trade-off* é particularmente relevante: ganhos de performance devem ser acompanhados por validações temporais e testes de robustez, de modo a evitar degradação fora da amostra e interpretações espúrias.

2.2 Generalização, *dataset shift* e *domain adaptation* em carteiras de crédito

A capacidade de um modelo manter desempenho quando aplicado fora do contexto em que foi treinado constitui um problema central em predição, frequentemente descrito como *dataset shift*. Esse fenômeno ocorre quando há alteração na distribuição das variáveis explicativas, na relação entre variáveis e desfecho, ou na composição amostral entre treinamento e aplicação, reduzindo a capacidade de generalização (Moreno-Torres et al., 2012; Webb et al., 2016). Em crédito corporativo, a ocorrência de *shift* é plausível mesmo sob conjuntos de variáveis padronizados, dado que carteiras podem diferir substantivamente em perfil de tomadores, condições contratuais, garantias e objetivos econômicos.

A literatura de *domain adaptation* e *transfer learning* propõe que modelos treinados em um domínio (carteira A) podem exigir ajustes para operar em um domínio distinto (carteira B), seja por reponderação amostral, incorporação de variáveis contextuais, reestimação parcial ou estratégias de integração de bases (Pan & Yang, 2010; Weiss et al., 2016). Em finanças, onde processos de validação regulatória são custosos e dados podem variar em disponibilidade e qualidade, abordagens que conciliem desempenho e estabilidade sob mudança de domínio tornam-se particularmente relevantes (Petropoulos et al., 2018; Markov et al., 2022).

No caso analisado, PRONAMPE e Capital de Giro representam domínios naturalmente distintos: além de objetivos e desenho institucional divergentes, há indícios de que as relações entre determinantes observáveis e risco de inadimplência podem não ser invariantes entre carteiras. Assim, a avaliação empírica da generalização cruzada opera como teste direto de robustez fora do domínio e como evidência indireta de heterogeneidade estrutural do risco, compatível com a noção de *dataset shift* em aplicações bancárias (Moreno-Torres et al., 2012; Webb et al., 2016).

2.3 Governança prudencial, estabilidade temporal e explicabilidade

Em ambientes regulados, modelos de risco devem atender simultaneamente a requisitos de desempenho, estabilidade e governança. No contexto brasileiro, a Resolução CMN nº 4.966/2021 estabelece princípios para classificação e mensuração associadas à identificação de ativos problemáticos, reforçando a necessidade de processos consistentes, documentados e monitoráveis ao longo do tempo (Banco Central do Brasil, 2021; Beerbaum, 2023). Nesse sentido, a robustez temporal (validação *out-of-time*), a prevenção de vazamento de informação

e a comparação estatística entre modelos tornam-se elementos essenciais para suportar decisões prudenciais e auditoráveis (Hurlin & Pérignon, 2023; Alonso & Escot, 2025).

Paralelamente, a adoção de ML amplia desafios de transparência, dado que modelos complexos tendem a operar como "caixas-pretas". Para mitigar esse problema, técnicas de *Explainable Artificial Intelligence* (XAI) permitem decompor previsões em contribuições atribuíveis às variáveis. Dentre essas técnicas, os valores SHAP (*SHapley Additive exPlanations*) oferecem propriedades desejáveis de consistência e aditividade local, viabilizando interpretação comparável entre algoritmos e análise global e local de importância (Lundberg & Lee, 2017; Bussmann et al., 2021). Em aplicações de risco, a explicabilidade assume papel duplo: (i) apoiar governança e justificativa técnica e (ii) permitir avaliar se determinantes de risco se mantêm estáveis entre segmentos, contribuindo para identificar heterogeneidade estrutural e potenciais mecanismos de *shift* (Marcinkevičs & Vogt, 2023; Fritz-Morgenthal et al., 2022).

2.4 Desenvolvimento das hipóteses

Com base nos fundamentos acima, este estudo organiza a investigação em três hipóteses, alinhadas a um desenho experimental que separa desempenho intracarteira, generalização cruzada e integração de bases. A formulação explicita a conexão entre (i) desempenho e estabilidade, (ii) heterogeneidade estrutural do risco e (iii) eficiência alocativa associada ao ranqueamento prudencial do risco.

H1 – Robustez preditiva intracarteiras.

Modelos de *Machine Learning* aplicados ao risco de crédito corporativo mantêm desempenho preditivo satisfatório e estabilidade estatística quando treinados e testados dentro de cada carteira (Capital de Giro e PRONAMPE).

A hipótese se apoia na evidência de que algoritmos de ML podem melhorar discriminação em *credit scoring* quando avaliados sob validações adequadas, especialmente em bases amplas (Lessmann et al., 2015), e requer, do ponto de vista prudencial, estabilidade temporal e consistência de métricas em avaliação fora do tempo.

H2 – Diferenças estruturais de risco e limites de transferibilidade.

A importância das variáveis preditoras e seus efeitos variam significativamente entre Capital de Giro e PRONAMPE, evidenciando heterogeneidade estrutural e limitações na generalização direta dos modelos entre carteiras.

Essa hipótese decorre da teoria de *dataset shift* e *domain adaptation*, segundo a qual mudanças de domínio podem alterar distribuições e relações preditivas, reduzindo o desempenho fora do domínio e modificando padrões de importância das variáveis (Moreno-Torres et al., 2012; Webb et al., 2016). A análise por SHAP permite operacionalizar essa comparação ao evidenciar diferenças sistemáticas na contribuição das variáveis para as previsões (Lundberg & Lee, 2017; Bussmann et al., 2021).

H3 – Eficiência alocativa do Modelo Combinado (Pooled).

A integração das carteiras em um Modelo Combinado (Pooled), com variável indicadora de origem, melhora a eficiência alocativa na priorização do risco (por exemplo, por medidas de ordenação como *Lift@10%* e KS), sem comprometer estabilidade e interpretabilidade.

A hipótese é consistente com a noção de que modelos treinados em bases ampliadas e heterogêneas podem capturar simultaneamente padrões compartilhados e específicos, potencialmente melhorando o ranqueamento de risco e a capacidade operacional de priorização, desde que respeitadas validações temporais e requisitos de governança (Bussmann et al., 2021; Hurlin & Pérignon, 2023).

3. Metodologia

3.1 Desenho da pesquisa e bases de dados

A pesquisa adota abordagem quantitativa, empírico-aplicada e não experimental, com foco na predição do risco de crédito corporativo por meio de técnicas de *Machine Learning*. O estudo utiliza dados históricos provenientes de uma instituição financeira brasileira classificada no segmento S1, abrangendo duas carteiras com características estruturais distintas: Capital de Giro e PRONAMPE.

As bases foram padronizadas para assegurar comparabilidade metodológica, resultando em um conjunto comum de 52 variáveis explicativas, contemplando informações cadastrais, contratuais e comportamentais observadas até o momento de referência. O período analisado

compreende operações ativas entre 2019 e 2025, totalizando aproximadamente 35,3 milhões de registros mensais e 1,17 milhão de contratos únicos, distribuídos entre as duas carteiras.

Para preservar a validade temporal e evitar vazamento de informação, todas as estimações respeitam separação *out-of-time*, com cortes temporais definidos antes do período de teste. Esse procedimento assegura que as previsões sejam realizadas exclusivamente com informações disponíveis no momento da decisão, em conformidade com boas práticas prudenciais e de governança de modelos.

A Tabela 1 sintetiza a caracterização das bases de dados utilizadas, incluindo período de referência, volume de registros e contratos, número de variáveis padronizadas e os respectivos cortes temporais *out-of-time*.

Tabela 1 – Caracterização das bases de dados

Carteira	Período de Referência	Registros Totais	Contratos Únicos	Variáveis Padronizadas	Corte Out-of-Time
Capital de Giro	01.2019 – 12.2024	19,2 milhões	646 mil	52	31.12.2021
PRONAMPE	06.2020 – 06.2025	16,1 milhões	525 mil	52	31.12.2023
Pooled (Global)	01.2019 – 06.2025	35,3 milhões	1,17 milhão	52	31.12.2021

Fonte: Elaboração própria.

3.2 Definição da variável-alvo

A variável dependente do estudo é a classificação de Ativo Problemático (AP), definida conforme os critérios estabelecidos na Resolução CMN nº 4.966/2021. Um contrato é classificado como AP quando apresenta evidências objetivas de deterioração de crédito, incluindo atraso relevante no pagamento ou outros indícios de perda esperada material.

A adoção dessa definição assegura alinhamento regulatório e confere relevância prática aos resultados, uma vez que a identificação de ativos problemáticos está diretamente associada à mensuração de perdas esperadas, à constituição de provisões e à alocação prudencial de capital.

3.3 Estratégia experimental

O desenho empírico é estruturado em três experimentos complementares, cada um orientado a testar uma hipótese específica do estudo:

i. **Experimento Baseline (H1)**

Os modelos são treinados e testados dentro da mesma carteira (Capital de Giro → Capital de Giro; PRONAMPE → PRONAMPE), respeitando a separação temporal *out-of-time*. Esse experimento avalia a **robustez preditiva intracarteiras** e a estabilidade estatística dos modelos.

ii. **Experimento de Generalização Cruzada – Cross (H2)**

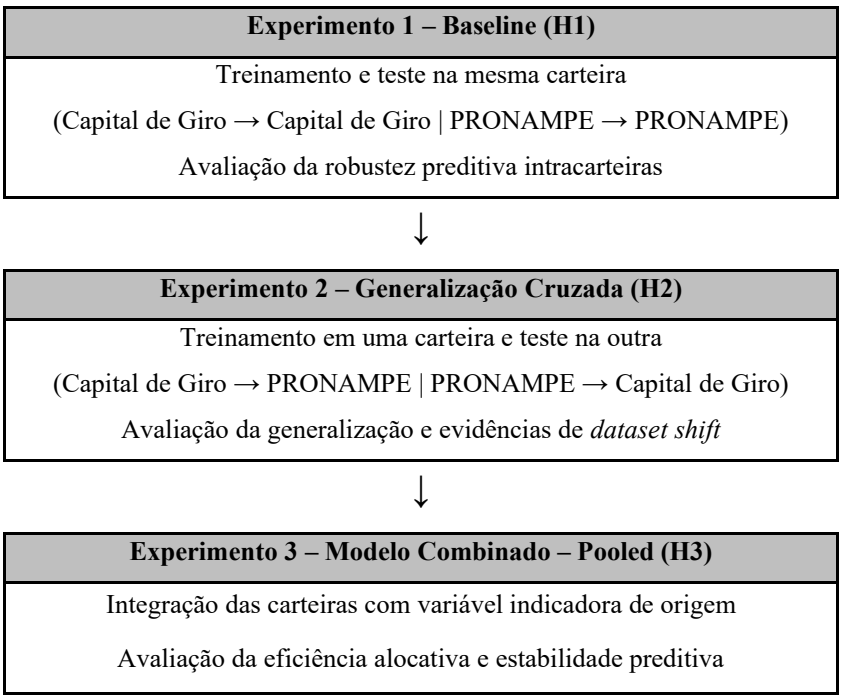
Os modelos treinados em uma carteira são aplicados diretamente na outra (Capital de Giro → PRONAMPE e PRONAMPE → Capital de Giro), sem reestimação de parâmetros. O objetivo é avaliar a **transferibilidade dos padrões de risco** e identificar evidências de *dataset shift* e heterogeneidade estrutural entre as carteiras.

iii. **Experimento de Modelo Combinado – Pooled (H3)**

As duas carteiras são integradas em uma única base consolidada, incluindo uma variável indicadora de origem da operação. Esse experimento investiga se a modelagem conjunta melhora a **eficiência alocativa** na priorização do risco, preservando estabilidade e interpretabilidade.

A Figura 1 sintetiza o desenho experimental adotado no estudo, ilustrando os três experimentos conduzidos e os respectivos fluxos de treinamento e teste associados às hipóteses H1, H2 e H3.

Figura 1 – Desenho experimental do estudo



Fonte: Elaboração própria (2026), a partir do pipeline de modelagem desenvolvido pelo autor.

A Tabela 2 sintetiza o delineamento experimental adotado, apresentando a descrição operacional de cada experimento, os respectivos cortes temporais e os objetivos analíticos associados às hipóteses H1, H2 e H3.

Tabela 2 – Estrutura geral e objetivos analíticos dos experimentos

Experimento	**Descrição Operacional**	**Corte Temporal (Treino/Teste)**	**Objetivo Analítico**
1 – Baseline	Treinamento e teste na mesma carteira.	Capital Giro: 31.12.2021 PRONAMPE: 31.12.2023	Avaliar robustez preditiva intracarteira e estabilidade temporal (**H1**).
2 – Cross	Treinamento em uma carteira e teste na outra.	Corte global: 31.12.2021	Testar capacidade de generalização intercarteiras e identificar ***dataset shift*** estrutural (**H2**).
3 – Pooled	Integração das carteiras com variável indicadora de origem.	Corte global: 31.12.2021	Avaliar ganhos de eficiência alocativa, estabilidade e integração de risco entre carteiras (**H3**).

Fonte: Elaboração própria.

3.4 Algoritmos de Machine Learning

Foram avaliados quatro algoritmos supervisionados amplamente utilizados na literatura de risco de crédito, selecionados por combinarem diferentes níveis de complexidade e interpretabilidade:

- Regressão Logística;
- Random Forest;
- Gradient Boosting Machine (GBM);
- Extreme Gradient Boosting (XGBoost).

A Regressão Logística é incluída como benchmark devido à sua transparência e ampla adoção regulatória, enquanto os métodos baseados em árvores e *boosting* representam abordagens não lineares com elevado poder discriminatório. A calibração dos modelos foi conduzida de forma a mitigar sobreajuste, respeitando validações temporais consistentes.

3.5 Métricas de avaliação e testes estatísticos

Considerando o desbalanceamento típico de eventos de inadimplência, o desempenho dos modelos foi avaliado por métricas adequadas a esse contexto, incluindo Recall, F1-score, AUC-PR, KS e Lift@10%. Essas métricas permitem avaliar tanto a capacidade discriminatória quanto a eficiência na priorização do risco, aspecto central para aplicações prudenciais.

A comparação entre modelos e experimentos foi conduzida por meio de testes estatísticos não paramétricos, apropriados para distribuições assimétricas e amostras emparelhadas. O teste de Wilcoxon pareado foi utilizado para comparar métricas de desempenho entre modelos e cenários; a correlação de Spearman e o teste de Mann–Whitney foram empregados na análise de diferenças estruturais de importância das variáveis; e o teste KS foi utilizado como medida adicional de separação entre distribuições de risco. Sempre que aplicável, adotaram-se correções para múltiplas comparações.

3.6 Interpretabilidade e análise de importância das variáveis

A interpretação dos modelos foi realizada por meio dos valores SHAP (*SHapley Additive exPlanations*), que decompõem as previsões em contribuições individuais das variáveis. Essa abordagem permite analisar a importância global dos atributos e comparar os

determinantes do risco entre as carteiras, fornecendo evidências sobre heterogeneidade estrutural e estabilidade dos padrões preditivos.

A análise comparativa dos valores SHAP entre Capital de Giro e PRONAMPE constitui elemento central para o teste da hipótese H2, ao permitir verificar se as variáveis exercem efeitos distintos sobre o risco de inadimplência em cada contexto.

Reprodutibilidade computacional

Todo o pipeline de preparação dos dados, estimação dos modelos, validação temporal, avaliação de desempenho e análise de interpretabilidade foi implementado em linguagem Python, utilizando bibliotecas consolidadas no ecossistema de *Machine Learning*. Os experimentos foram executados em ambiente Google Colab, com controle explícito das etapas do pipeline e registro estruturado das saídas intermediárias e finais. Com vistas à transparência metodológica e à reprodutibilidade computacional, o notebook completo contendo o código-fonte utilizado na execução das análises encontra-se publicamente disponível em ambiente Google Colab, no endereço: https://doi.org/10.5281/zenodo.19035323

4. Resultados

4.1 Robustez preditiva intracarteiras – Experimento Baseline (H1)

O primeiro conjunto de resultados avalia o desempenho dos modelos de *Machine Learning* quando treinados e testados dentro da mesma carteira de crédito, com separação temporal *out-of-time*. Os resultados indicam que, em ambas as carteiras, os modelos não lineares apresentam desempenho consistente e estatisticamente superior ao benchmark linear, especialmente quando avaliados por métricas sensíveis ao desbalanceamento, como Recall, F1-score e AUC-PR. Observa-se, adicionalmente, estabilidade das métricas entre janelas temporais, sugerindo ausência de sobreajuste relevante e adequação do desenho de validação adotado.

Embora os níveis absolutos de desempenho difiram entre Capital de Giro e PRONAMPE, os modelos demonstram capacidade de discriminação satisfatória dentro de cada contexto, corroborando a hipótese de que algoritmos de ML são capazes de capturar padrões de risco específicos quando aplicados em ambientes homogêneos do ponto de vista estrutural.

A Tabela 3 apresenta o desempenho dos modelos no experimento Baseline para as carteiras de Capital de Giro e PRONAMPE, com base nas métricas AUC, F1-score e Recall, calculadas em amostras *out-of-time*. As diferenças de desempenho entre os algoritmos foram avaliadas por meio de testes estatísticos não paramétricos.

Tabela 3 – Resultados Baseline (Capital de Giro + PRONAMPE)

Experimento	Modelo	AUC	F1	Recall	Precisao	Acurácia
Baseline Capital Giro	GB	0.9980	0.9818	0.9921	0.9717	0.9749
	XGB	0.9980	0.9811	0.9857	0.9765	0.9740
	RF	0.9976	0.9808	0.9872	0.9745	0.9735
	LR	0.9942	0.9669	0.9521	0.9822	0.9554
Baseline PRONAMPE	LR	0.9514	0.5462	0.9983	0.3760	0.8892
	XGB	0.9504	0.5440	0.6230	0.4827	0.9302
	RF	0.9308	0.3831	0.2534	0.7849	0.9455
	GB	0.9600	0.3678	0.2671	0.5905	0.9387

Fonte: Elaboração própria.

Em conjunto, esses achados fornecem evidência empírica favorável à H1 – Robustez preditiva intracarteiras, ao indicar que os modelos mantêm desempenho e estabilidade quando aplicados no mesmo domínio de crédito.

4.2 Generalização cruzada e heterogeneidade estrutural - Experimento Cross (H2)

O segundo experimento avalia a capacidade de generalização dos modelos quando aplicados fora do domínio em que foram treinados, por meio da generalização cruzada entre Capital de Giro e PRONAMPE. Esse teste é central para investigar a existência de *dataset shift* e diferenças estruturais de risco, conforme a hipótese H2.

Os resultados evidenciam deterioração significativa do desempenho quando os modelos treinados em uma carteira são aplicados diretamente na outra, sem reestimação de parâmetros. Essa perda é observada de forma consistente nas métricas de discriminação e priorização, indicando que os padrões de risco aprendidos não são plenamente transferíveis entre os dois segmentos. Testes estatísticos não paramétricos confirmam que as diferenças de desempenho em relação aos cenários Baseline são estatisticamente significativas.

Esses achados sugerem que, embora as bases compartilhem o mesmo conjunto de variáveis explicativas, as relações entre essas variáveis e o risco de inadimplência diferem entre as carteiras, refletindo heterogeneidade estrutural compatível com a literatura de *domain adaptation*.

A Tabela 4 sintetiza os resultados do experimento de generalização cruzada, apresentando o desempenho dos modelos nos cenários Capital de Giro → PRONAMPE e PRONAMPE → Capital de Giro.

Tabela 4 – Resultados Cross (CG → PRONAMPE; PRONAMPE → CG)

Experimento	Modelo	AUC	F1	Recall	Precisao	Acurácia
Cross CG > PRONAMPE	GB	0.9717	0.5171	0.6821	0.4164	0.9553
	RF	0.9722	0.4765	0.5768	0.4059	0.9556
	XGB	0.9714	0.4391	0.4722	0.4104	0.9577
	LR	0.9627	0.4028	0.9956	0.2524	0.8964
Cross PRONAMPE > CG	LR	0.9805	0.8006	1.000	0.6676	0.6876
	GB	0.9588	0.7134	0.5647	0.9684	0.7153
	XGB	0.8256	0.5453	0.3861	0.9279	0.5961
	RF	0.9594	0.1307	0.0701	0.9730	0.4154

Fonte: Elaboração própria.

Para aprofundar a análise estrutural, a importância das variáveis foi examinada por meio dos valores SHAP estimados nos modelos Baseline de cada carteira. A Tabela 5 apresenta as cinco variáveis mais relevantes segundo a média do valor absoluto dos SHAP values, evidenciando que, embora exista um núcleo comum de variáveis relevantes, a magnitude e o papel relativo de seus efeitos diferem substancialmente entre Capital de Giro e PRONAMPE. Esses resultados reforçam a interpretação de heterogeneidade estrutural nos determinantes do risco de crédito, consistente com a deterioração observada no experimento de generalização cruzada.

Tabela 5 – Principais variáveis segundo a importância SHAP nos modelos Baseline (Top-5 por carteira)

Rank	Variável	SHAP – CG	SHAP – PRONAMPE
1	vr_divida_vincenda	3.3646	0.1744
2	idade_contrato_dias	1.9699	0.0154
3	prazo_remanescente_dias	0.9844	0.0220
4	valor_base_calculo	0.3010	0.0036
5	prazo_contrato_dias	0.1046	0.0051

Fonte: Elaboração própria.

Nota: Valores correspondem à média do valor absoluto dos SHAP *values* da classe positiva no conjunto de teste *out-of-time.*

Dessa forma, a Hipótese H2 é parcialmente confirmada: há capacidade mensurável de generalização, mas limitada por diferenças estruturais de risco que exigem calibragem específica por carteira ou abordagens integradas, em linha com a literatura de *domain adaptation* em risco de crédito (Liu et al., 2024; Suryanto et al., 2022; Maldonado et al., 2021; Penikas, 2020).

4.3 Eficiência alocativa do modelo combinado – Experimento Pooled (H3)

O terceiro experimento investiga se a integração das carteiras em um Modelo Combinado (Pooled), com variável indicadora de origem, resulta em ganhos de eficiência alocativa na priorização do risco. O foco recai sobre métricas de ordenação, como Lift@10% e KS, relevantes para aplicações prudenciais e operacionais.

Os resultados indicam que o modelo combinado apresenta desempenho superior na priorização do risco em relação aos modelos estimados separadamente, especialmente nos decis superiores da distribuição de score. Esses ganhos são estatisticamente significativos e refletem maior capacidade do modelo em concentrar eventos de inadimplência nas faixas superiores de risco, aspecto central para estratégias de monitoramento e alocação prudencial de capital.

Importa destacar que o ganho de eficiência não implica eliminação das diferenças estruturais observadas entre as carteiras. Ao contrário, os resultados sugerem que o modelo Pooled se beneficia da integração de informações complementares, preservando a heterogeneidade por meio da variável indicadora de origem.

A Tabela 6 apresenta os modelos vencedores em cada experimento, com ênfase nas métricas de eficiência alocativa, como KS e Lift@10%. A comparação dos resultados indica que o Modelo Combinado (Pooled) apresenta ganhos relevantes de eficiência alocativa na priorização do risco quando comparado ao modelo Baseline de Capital de Giro, especialmente nos decis superiores da distribuição de score. Embora não supere o melhor desempenho observado em modelos segmentados específicos, o modelo Pooled oferece uma solução integrada com capacidade consistente de ranqueamento do risco em ambientes com múltiplas carteiras.

Tabela 6 – Modelos vencedores por experimento e métricas de eficiência alocativa

Experimento	Modelo	F1	AUC	Recall	Precisao	Acurácia	Gini	Lift@10%
Baseline Capital Giro	GB	0.9818	0.9980	0.9921	0.9717	0.9749	0.9959	1.4607
Baseline PRONAMPE	LR	0.5462	0.9514	0.9983	0.3760	0.8892	0.9027	6.4639
Cross CG > PRONAMPE	GB	0.5171	0.9717	0.6821	0.4164	0.9553	0.9434	9.9743
Cross PRONAMPE > CG	LR	0.8006	0.9805	1.0000	0.6676	0.6876	0.9611	1.5803
Modelo Pooled	RF	0.9277	0.9946	0.9023	0.9545	0.9620	0.9892	3.7010

Fonte: Elaboração própria.

Esses achados fornecem evidência favorável à H3 – Eficiência alocativa do modelo combinado, ao indicar que a modelagem integrada pode ampliar a priorização do risco em contextos de múltiplas carteiras, sem eliminar a necessidade de abordagens segmentadas quando o objetivo é maximizar desempenho específico.

4.4 Síntese dos resultados empíricos

Em síntese, os resultados empíricos confirmam que os modelos de *Machine Learning* apresentam desempenho robusto e estável quando treinados e testados dentro do mesmo domínio de crédito, em validações temporais rigorosas *out-of-time* (H1). Em contraste, a aplicação cruzada entre as carteiras de Capital de Giro e PRONAMPE evidencia deterioração significativa do desempenho, refletindo heterogeneidade estrutural nos determinantes do risco, mesmo sob bases de variáveis padronizadas (H2). Por fim, a modelagem combinada (Pooled) demonstra ganhos consistentes de eficiência alocativa na priorização do risco, ao concentrar

eventos de inadimplência nos decis superiores da distribuição de score, sem eliminar as diferenças estruturais entre os segmentos (H3).

Esses achados estabelecem uma conexão direta entre o desenho experimental adotado e as hipóteses formuladas, evidenciando os limites da generalização direta entre carteiras heterogêneas e o papel da modelagem integrada como alternativa pragmática sob uma perspectiva prudencial. A seção seguinte discute as implicações teóricas, empíricas e prudenciais desses resultados à luz da literatura sobre modelagem preditiva, *domain adaptation* e governança do risco de crédito.

5. Discussão

Os resultados empíricos obtidos neste estudo oferecem evidências consistentes sobre os limites e as potencialidades da aplicação de modelos de *Machine Learning* ao risco de crédito corporativo em carteiras estruturalmente distintas. A discussão a seguir articula esses achados às hipóteses formuladas e à literatura sobre modelagem preditiva, *domain adaptation* e governança prudencial.

Em relação à hipótese H1, os resultados do experimento Baseline confirmam que modelos de *Machine Learning* apresentam desempenho preditivo robusto quando aplicados dentro do domínio para o qual foram estimados. A estabilidade observada sob validações temporais rigorosas reforça evidências anteriores de que algoritmos não lineares podem superar abordagens tradicionais em ambientes de crédito corporativo, desde que avaliados com métricas adequadas a bases desbalanceadas e com separação *out-of-time* (Lessmann et al., 2015; Zou & Gao, 2022). Do ponto de vista prudencial, esses achados indicam que modelos de ML podem ser incorporados como instrumentos de monitoramento e alerta precoce, desde que inseridos em arcabouços de governança que priorizem estabilidade e validação contínua.

Por outro lado, os resultados do experimento de generalização cruzada oferecem evidências claras em favor da hipótese H2, ao demonstrar deterioração significativa do desempenho quando os modelos são aplicados fora do domínio de treinamento. Essa perda de capacidade discriminatória, observada de forma consistente entre métricas e algoritmos, é compatível com a literatura de *dataset shift* e reforça a noção de que padrões de risco não são invariantes entre carteiras com desenhos institucionais e perfis de tomadores distintos (Moreno-Torres et al., 2012; Webb et al., 2016). Mesmo sob um conjunto padronizado de variáveis, as relações entre atributos observáveis e inadimplência mostraram-se dependentes do contexto da carteira, limitando a transferibilidade direta dos modelos.

A análise de explicabilidade via valores SHAP aprofunda essa interpretação ao revelar diferenças sistemáticas na importância e no efeito das variáveis entre Capital de Giro e PRONAMPE. A baixa correlação entre rankings de importância e as diferenças estatisticamente significativas nos valores médios de |SHAP| sugerem heterogeneidade estrutural nos determinantes do risco, corroborando evidências de que políticas públicas de crédito, garantias e critérios de elegibilidade influenciam os mecanismos geradores de inadimplência. Esses achados reforçam o papel da explicabilidade não apenas como requisito regulatório, mas como ferramenta analítica para identificação de *shift* estrutural entre segmentos (Bussmann et al., 2021; Marcinkevičs & Vogt, 2023).

No que se refere à hipótese H3, os resultados do modelo combinado indicam ganhos consistentes de eficiência alocativa na priorização do risco, especialmente em métricas sensíveis à ordenação, como Lift@10% e KS. Esses ganhos sugerem que a integração das carteiras permite ao modelo capturar simultaneamente padrões compartilhados e específicos, ampliando a capacidade de ranqueamento do risco nos decis superiores da distribuição. Importa destacar, contudo, que essa melhora não implica homogeneização estrutural entre as carteiras. Ao contrário, a evidência sugere que a variável indicadora de origem desempenha papel relevante ao permitir que o modelo diferencie contextos, mitigando parcialmente os efeitos adversos da heterogeneidade.

Sob a ótica da gestão prudencial, esse resultado possui implicações relevantes. Enquanto a generalização direta entre carteiras se mostrou limitada, a modelagem integrada emerge como alternativa viável para apoiar estratégias de priorização e monitoramento do risco em ambientes com múltiplos segmentos. Essa abordagem é particularmente útil em contextos operacionais nos quais decisões são orientadas por ranqueamento relativo, como alocação de esforços de acompanhamento, definição de limites e estratégias de provisão prospectiva. Ao mesmo tempo, os resultados reforçam que modelos combinados não substituem a necessidade de análises segmentadas quando o objetivo é compreender os determinantes específicos do risco.

De forma mais ampla, os achados contribuem para a literatura de *domain adaptation* em finanças ao fornecer evidência empírica em larga escala sobre os limites da transferibilidade de modelos de crédito corporativo entre domínios heterogêneos. Diferentemente de abordagens que assumem invariância estrutural, os resultados indicam que a estabilidade preditiva deve ser avaliada explicitamente como hipótese empírica, e não presumida. Nesse sentido, o estudo reforça a importância de desenhos experimentais que separem desempenho intracarteiras,

generalização cruzada e integração de bases, especialmente em ambientes regulados nos quais decisões baseadas em modelos possuem implicações sistêmicas.

Em síntese, a discussão dos resultados sugere que o uso de *Machine Learning* no risco de crédito corporativo deve ser orientado por uma lógica de complementaridade: modelos segmentados oferecem maior fidelidade estrutural, enquanto modelos integrados podem ampliar a eficiência alocativa sob determinadas aplicações prudenciais. A combinação dessas abordagens, apoiada por mecanismos de explicabilidade e validação estatística, contribui para uma gestão de risco mais robusta, transparente e alinhada às exigências regulatórias.

6. Conclusão

Este estudo investigou a aplicação de modelos de *Machine Learning* à predição do risco de crédito corporativo em carteiras com características estruturais distintas, especificamente Capital de Giro e PRONAMPE, com o objetivo de avaliar robustez preditiva, capacidade de generalização e eficiência alocativa sob uma perspectiva prudencial. A partir de um desenho experimental estruturado em três experimentos — Baseline, Generalização Cruzada e Modelo Combinado (Pooled) —, os resultados fornecem evidências empíricas relevantes para a literatura e para a prática de gestão do risco de crédito em ambientes regulados.

Os achados indicam que os modelos de *Machine Learning* apresentam desempenho robusto e estabilidade estatística quando aplicados dentro de cada carteira, corroborando a hipótese de robustez preditiva intracarteiras (H1). Esse resultado reforça a viabilidade do uso de algoritmos não lineares como instrumentos de monitoramento e apoio à gestão prudencial, desde que submetidos a validações temporais rigorosas e métricas adequadas a bases desbalanceadas.

Por outro lado, a deterioração significativa observada nos experimentos de generalização cruzada evidencia limites claros à transferibilidade direta dos modelos entre carteiras heterogêneas, confirmando a hipótese de diferenças estruturais de risco (H2). A análise de explicabilidade via valores SHAP reforça essa interpretação ao demonstrar que a importância e o efeito das variáveis diferem substancialmente entre Capital de Giro e PRONAMPE, mesmo sob um conjunto padronizado de atributos. Esses resultados contribuem para a literatura de *dataset shift* e *domain adaptation* em finanças, ao evidenciar empiricamente que a generalização deve ser tratada como hipótese a ser testada, e não como propriedade presumida dos modelos.

Adicionalmente, os resultados do modelo combinado indicam ganhos consistentes de eficiência alocativa na priorização do risco, especialmente em métricas de ordenação relevantes para aplicações prudenciais, como Lift@10% e KS, oferecendo suporte à hipótese H3. Esses ganhos sugerem que a integração de carteiras pode capturar informações complementares e ampliar a capacidade de ranqueamento do risco, sem eliminar a heterogeneidade estrutural entre os segmentos. Do ponto de vista prático, isso indica que modelos combinados podem ser úteis em contextos operacionais orientados à priorização e ao monitoramento, desde que complementados por análises segmentadas quando o objetivo for compreender os determinantes específicos do risco.

Apesar das contribuições, o estudo apresenta limitações que devem ser consideradas. Os resultados são baseados em dados de uma única instituição financeira brasileira classificada no segmento S1, o que pode limitar a generalização externa para outros contextos institucionais. Além disso, a análise concentrou-se em algoritmos supervisionados amplamente utilizados, não contemplando abordagens mais recentes de *transfer learning* explícito ou adaptação dinâmica entre domínios, que podem oferecer ganhos adicionais de generalização.

Como agenda de pesquisa futura, estudos podem explorar técnicas formais de *domain adaptation*, como reponderação amostral, aprendizado multi-tarefa ou modelos hierárquicos, bem como avaliar a estabilidade dos resultados sob choques macroeconômicos ou mudanças regulatórias. Adicionalmente, a ampliação da análise para múltiplas instituições e outros segmentos de crédito pode aprofundar a compreensão sobre a invariância, ou não, dos determinantes do risco em ambientes heterogêneos.

Em síntese, o estudo demonstra que o uso de *Machine Learning* no risco de crédito corporativo deve ser orientado por uma lógica de complementaridade entre modelos segmentados e integrados, apoiada por validação estatística rigorosa e mecanismos de explicabilidade. Essa abordagem contribui para uma gestão de risco mais robusta, transparente e alinhada às exigências prudenciais, ao mesmo tempo em que avança a compreensão acadêmica sobre generalização e heterogeneidade estrutural em modelos preditivos de crédito.

Referências

Alonso, J. V., & Escot, L. (2025). Robust cross-validation of predictive models used in credit default risk. *Applied Sciences, 15*(10), 5495. https://doi.org/10.3390/app15105495

Araujo, R. F. de, Alves, V. L. de S., Silva, N. G. da, Monteiro, J. G. M. A., Palludeto, A. W. A., & Borghi, R. A. Z. (2021). Medidas fiscais e parafiscais diante da pandemia de Covid-19: experiências internacionais selecionadas. *Revista Tempo do Mundo, 26*, 35–75. http://dx.doi.org/10.38116/rtm26art1

Banco Central do Brasil. (2021). *Resolução CMN nº 4.966, de 25 de agosto de 2021*. https://www.bcb.gov.br/normas/exibe/res/4966

Beerbaum, D. (2023). The new impairment loss credit loss model under IFRS 9 – Post transition model implications. *SSRN Electronic Journal*. https://ssrn.com/abstract=4750458

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics, 57*, 203–216. https://doi.org/10.1007/s10614-020-10042-0

Dong, G., Lai, K. K., & Yen, J. (2012). Credit scorecard based on logistic regression with random coefficients. *Procedia Computer Science, 1*, 2463–2468. https://doi.org/10.1016/j.procs.2010.04.278

Fritz-Morgenthal, S., Hein, B., & Papenbrock, J. (2022). Financial Risk Management and Explainable, Trustworthy, Responsible AI. *Frontiers in Artificial Intelligence, 5*, 779799. https://doi.org/10.3389/frai.2022.779799

Hurlin, C., & Pérignon, C. (2023). Machine learning and IRB capital requirements: Advantages, risks, and recommendations. *SSRN Electronic Journal*. https://ssrn.com/abstract=4483793

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring. *European Journal of Operational Research, 247*(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Li, Y., Stasinakis, C., & Yeo, W. M. (2022). A hybrid XGBoost-MLP model for credit risk assessment on digital supply chain finance. *Forecasting, 4*(1), 184–207. https://doi.org/10.3390/forecast4010011

Liu, Z., Zhang, G., & Lu, J. (2024). Semi-supervised heterogeneous domain adaptation for few-sample credit risk classification. *Neurocomputing, 596*, 127948. https://doi.org/10.1016/j.neucom.2024.127948

Louzis, D. P., Vouldis, A. T., & Metaxas, V. L. (2012). Macroeconomic and bank-specific determinants of non-performing loans. *Journal of Banking & Finance, 36*(4), 1012–1027. https://doi.org/10.1016/j.jbankfin.2011.10.012

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

Maldonado, S., López, J., & Iturriaga, A. (2021). Out-of-time cross-validation strategies for classification in the presence of dataset shift. *Applied Intelligence*. https://doi.org/10.1007/s10489-021-02735-2

Marcinkevičs, R., & Vogt, J. E. (2023). Interpretability and explainability: A machine learning zoo mini-tour. *WIREs Data Mining and Knowledge Discovery*, e1493. https://doi.org/10.48550/arXiv.2012.01805

Markov, A., Seleznyova, Z., & Lapshin, V. (2022). Credit scoring methods: Latest trends and points to consider. *The Journal of Finance and Data Science, 8*(3), 180–201. https://doi.org/10.1016/j.jfds.2022.07.002

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., & Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern Recognition, 45*(1), 521–530. https://doi.org/10.1016/j.patcog.2011.06.019

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Penikas, H. (2020). History of the Basel internal-ratings-based (IRB) credit risk regulation. *Model Assisted Statistics and Applications, 15*(1), 81–98. https://doi.org/10.3233/MAS-190480

Petropoulos, A., Siakoulis, V., Stavroulakis, E., & Klamargias, A. (2018). A robust machine learning approach for credit risk analysis. *Bank for International Settlements Conference Proceedings*. https://www.bis.org/ifc/publ/ifcb49_49.pdf

Suryanto, H., Mahidadia, A., Bain, M., Guan, C., & Guan, A. (2022). Credit risk modeling using transfer learning and domain adaptation. *Frontiers in Artificial Intelligence, 5*, 868232. https://doi.org/10.3389/frai.2022.868232

Ustarz, Y., & Fanta, A. B. (2021). Financial development and economic growth in sub-Saharan Africa: A sectoral perspective. *Cogent Economics & Finance, 9*(1), 1934976. https://doi.org/10.1080/23322039.2021.1934976

Vasconcellos de Paula, D. A., Artes, R., Ayres, F., & Fonseca Minardi, A. M. A. (2019). Estimating credit and profit scoring of a Brazilian credit union with logistic regression and machine-learning techniques. *RAUSP Management Journal, 54*(3), 321–336. https://doi.org/10.1108/RAUSP-03-2018-0003

Webb, G. I., Hyde, R., Cao, H., Nguyen, H. L., & Petitjean, F. (2016). Characterizing concept drift. *Data Mining and Knowledge Discovery, 30*(4), 964–994. https://doi.org/10.1007/s10618-015-0448-4

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data, 3*(1), 9. https://doi.org/10.1186/s40537-016-0043-6

Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. (2019). A study on predicting loan default based on the random forest algorithm. *Procedia Computer Science, 162*, 503–513. https://doi.org/10.1016/j.procs.2019.12.017

Zou, Y., & Gao, C. (2022). Extreme learning machine enhanced gradient boosting for credit scoring. *Algorithms, 15*(5), 149. https://doi.org/10.3390/a15050149