RAFAEL ALMEIDA PEREIRA MELO

PREDIÇÃO DE RISCO DE CRÉDITO COM APRENDIZADO DE MÁQUINA SUPERVISIONADO: UM ESTUDO DE CASO COM DADOS DESBALANCEADOS

LAVRAS – MG

2026

RAFAEL ALMEIDA PEREIRA MELO

PREDIÇÃO DE RISCO DE CRÉDITO COM APRENDIZADO DE MÁQUINA **SUPERVISIONADO**: UM ESTUDO DE CASO COM DADOS DESBALANCEADOS

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

Prof. Dr. Paulo Henrique Sales Guimarães
Orientador

Prof. Dr. Marcel Irving Pereira Melo
Coorientador

LAVRAS – MG
2026

Ficha Catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Melo, Rafael Almeida Pereira.
Predição de risco de crédito com aprendizado de máquina supervisionado : um estudo de caso com dados desbalanceados / Rafael Almeida Pereira Melo. - 2025.
69 p. : il.

Orientador: Paulo Henrique Sales Guimarães
Coorientador: Marcel Irving Pereira Melo

Dissertação (Mestrado Acadêmico) - Universidade Federal de Lavras, 2025.
Bibliografia.

1. Regressão logística. 2. Inadimplência. 3. Classificação de dados. I. Guimarães, Paulo Henrique Sales. II. Melo, Marcel Irving Pereira. III. Universidade Federal de Lavras. IV. Título.

RAFAEL ALMEIDA PEREIRA MELO

PREDIÇÃO DE RISCO DE CRÉDITO COM APRENDIZADO DE MÁQUINA SUPERVISIONADO: UM ESTUDO DE CASO COM DADOS DESBALANCEADOS

CREDIT RISK PREDICTION WITH SUPERVISED MACHINE LEARNING: A CASE STUDY WITH IMBALANCED DATA

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

APROVADA em 30 de setembro de 2025.
Prof. Darcy Ramos da Silva Neto
Prof. Luiz Otavio de Oliveira Pala
Prof. Manoel Vitor de Souza Veloso

gov.br
Documento assinado digitalmente
PAULO HENRIQUE SALES GUIMARAES
Data: 23/01/2026 15:43:13-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Paulo Henrique Sales Guimarães
Orientador

Prof. Dr. Marcel Irving Pereira Melo
Co-Orientador

LAVRAS – MG
2026

AGRADECIMENTOS

Agradeço primeiramente ao meu orientador, Paulo Henrique Sales Guimarães, pela orientação, dedicação e apoio ao longo do desenvolvimento deste trabalho.

Ao meu coorientador, Marcel Irving Pereira Melo, pela colaboração e pelas valiosas contribuições que enriqueceram esta pesquisa.

À minha família e amigos, pelo incentivo constante e pela compreensão nos momentos de maior dedicação acadêmica.

Agradeço ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio financeiro por meio da bolsa concedida, sem o qual este trabalho não teria sido possível.

Por fim, agradeço a todos que, direta ou indiretamente, contribuíram para a realização desta dissertação.

RESUMO

Este trabalho avalia a aplicação da Regressão Logística penalizada via LASSO na predição de risco de crédito, com foco no tratamento de dados desbalanceados, uma característica comum em bases financeiras, onde a proporção de inadimplentes é geralmente baixa. Foram utilizadas 36 bases sintéticas, geradas por simulação, com diferentes combinações de tamanho amostral e proporção de inadimplentes, e o modelo foi calibrado por meio de validação cruzada e ajuste dinâmico do *cut-off* (ponto de corte ou limiar de decisão), sem uso de técnicas de reamostragem como SMOTE *(Synthetic Minority Over-sampling Technique*, que gera exemplos sintéticos da classe minoritária) ou *undersampling* (redução da classe majoritária para equilibrar a distribuição), tendo sido adotados procedimentos de reamostragem com reposição da base original para gerar diferentes cenários de tamanho amostral e desbalanceamento. Os resultados mostraram desempenho robusto em métricas como AUC, F1 Score e acurácia balanceada, mesmo em cenários com forte desbalanceamento, isto é, quando a proporção de inadimplentes é extremamente baixa (entre 1% e 5%). A penalização LASSO contribuiu para a seleção automática de variáveis relevantes, mantendo a interpretabilidade do modelo. Como contribuição prática, o estudo demonstra que, com ajustes criteriosos, é possível obter modelos estatísticos simples, eficazes e compatíveis com exigências regulatórias, oferecendo suporte confiável à gestão de risco de crédito. Além disso, os achados podem ser generalizados para outros contextos que envolvam dados desbalanceados, ampliando a aplicabilidade da abordagem proposta.

Palavras-chave: regressão logística; regularização; inadimplência; classificação de dados.

ABSTRACT

This work evaluates the application of penalized Logistic Regression via LASSO in credit risk prediction, focusing on the treatment of imbalanced data, a common characteristic in financial datasets where the proportion of defaulters is usually low. A total of 36 synthetic datasets were generated through simulation, with different combinations of sample size and proportion of defaulters. The model was calibrated using cross-validation and dynamic adjustment of the *cut-off* (decision threshold), without employing resampling techniques such as SMOTE *(Synthetic Minority Over-sampling Technique*, which generates synthetic examples of the minority class) or *undersampling* (reducing the majority class to balance the distribution). Instead, resampling procedures with replacement from the original dataset were adopted to generate different scenarios of sample size and imbalance. The results showed robust performance in metrics such as AUC, F1 Score, and balanced accuracy, even in scenarios with severe imbalance, that is, when the proportion of defaulters is extremely low (between 1% and 5%). The LASSO penalization contributed to the automatic selection of relevant variables, while maintaining model interpretability. As a practical contribution, the study demonstrates that, with careful adjustments, it is possible to obtain statistical models that are simple, effective, and compliant with regulatory requirements, providing reliable support for credit risk management. Furthermore, the findings can be generalized to other contexts involving imbalanced data, expanding the applicability of the proposed approach.

Keywords: logistic regression; regularization; credit default; data classification.

INDICADORES DE IMPACTO

Este trabalho contribui para o avanço metodológico na avaliação de modelos de classificação em cenários de desbalanceamento de classes, sem recorrer a técnicas artificiais de oversampling como o SMOTE. A adoção de procedimentos de reamostragem com reposição, controlando o tamanho amostral e a proporção da classe positiva, permite compreender de forma mais precisa os limites e a robustez dos modelos em situações que refletem a realidade dos dados financeiros.

Do ponto de vista científico e tecnológico, os resultados oferecem subsídios para pesquisas futuras em aprendizado de máquina aplicado a dados desbalanceados, destacando a importância da calibragem dinâmica do *cut-off* e da escolha criteriosa das métricas de avaliação.

Sob a perspectiva social e regulatória, os resultados reforçam a necessidade de políticas de concessão de crédito baseadas em evidências, capazes de equilibrar inclusão financeira e mitigação de risco. O público diretamente beneficiado são instituições financeiras e seus clientes, especialmente em territórios com alta vulnerabilidade socioeconômica.

No âmbito cultural e acadêmico, este trabalho estimula a adoção de práticas estatísticas mais rigorosas e contextualizadas, contribuindo para a formação de pesquisadores e profissionais conscientes das limitações e implicações práticas dos modelos de classificação.

Por fim, sob a ótica econômica e operacional, a compreensão da estabilidade das métricas em diferentes tamanhos amostrais favorece a redução de custos associados a decisões equivocadas de crédito. Modelos mais robustos e calibrados reduzem perdas financeiras por inadimplência e minimizam gastos com revisões manuais ou provisões excessivas, caracterizando impacto direto na eficiência das instituições financeiras.

Em síntese, os impactos deste trabalho se manifestam em diferentes dimensões, científica, social, cultural e econômica, e contribuem tanto para o avanço metodológico quanto para a aplicação prática em políticas de crédito, alinhando-se às áreas temáticas de Tecnologia e Produção e Educação da Política Nacional de Extensão, além de dialogar com os Objetivos de Desenvolvimento Sustentável da ONU.

IMPACT INDICATORS

This work contributes to the methodological advancement in the evaluation of classification models in scenarios of class imbalance, without resorting to artificial oversampling techniques such as SMOTE. The adoption of resampling procedures with replacement, controlling the sample size and the proportion of the positive class, allows for a more precise understanding of the limits and robustness of the models in situations that reflect the reality of financial data.

From a scientific and technological perspective, the results provide support for future research in machine learning applied to imbalanced data, highlighting the importance of dynamic calibration of the *cut-off* and the careful choice of evaluation metrics.

From a social and regulatory perspective, the results reinforce the need for evidence-based credit granting policies, capable of balancing financial inclusion and risk mitigation. The directly benefited public includes financial institutions and their clients, especially in territories with high socioeconomic vulnerability.

In the cultural and academic sphere, this work encourages the adoption of more rigorous and contextualized statistical practices, contributing to the training of researchers and professionals who are aware of the limitations and practical implications of classification models.

Finally, from an economic and operational standpoint, understanding the stability of metrics across different sample sizes favors the reduction of costs associated with misguided credit decisions. More robust and calibrated models reduce financial losses due to default and minimize expenses with manual reviews or excessive provisions, representing a direct impact on the efficiency of financial institutions.

In summary, the impacts of this work manifest in different dimensions, scientific, social, cultural, and economic, and contribute both to methodological advancement and to practical application in credit policies, aligning with the thematic areas of Technology and Production and Education of the National Extension Policy, as well as dialoguing with the United Nations Sustainable Development Goals.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

Em um mundo globalizado, o crédito desempenha um papel relevante no consumo e pode influenciar o crescimento econômico e a redução das desigualdades sociais. Ao facilitar o acesso a bens, serviços, educação e empreendedorismo, o crédito permite que indivíduos e empresas ampliem sua capacidade produtiva e melhorem suas condições de vida. Quando distribuído de forma responsável e inclusiva, ele atua como instrumento de mobilidade social, especialmente para populações com menor renda ou acesso limitado a recursos financeiros. No entanto, a crescente complexidade das operações financeiras e a dependência de tecnologias tornam os riscos associados à inadimplência cada vez mais desafiadores de prever e gerenciar, exigindo modelos analíticos mais robustos e capazes de lidar com grandes volumes de dados. A aplicação de técnicas avançadas de modelagem, como redes neurais, regressões penalizadas e algoritmos híbridos, tem se mostrado eficaz na previsão da inadimplência, especialmente em contextos com múltiplas variáveis econômicas e alta volatilidade (Farias, 2023).

A pandemia de COVID-19 agravou esses desafios ao ameaçar as condições de crédito de indivíduos, empresas e países. A redução da renda potencial, o fechamento de negócios e os altos níveis de endividamento resultaram em uma crise de solvência para diversas empresas, ameaçando a estabilidade econômica global. Uma crise de solvência ocorre quando uma instituição, seja uma empresa ou um governo, enfrenta dificuldades significativas para honrar suas dívidas devido à insuficiência de ativos ou receitas em relação às obrigações financeiras. Assim, as instituições financeiras enfrentaram restrições de liquidez e adotaram posturas mais cautelosas em relação ao risco de crédito, evidenciando a necessidade de ferramentas mais robustas e precisas para assegurar a saúde do sistema financeiro (Yin; Han; Wong, 2022).

Atualmente, as instituições financeiras lidam com grandes volumes de dados heterogêneos, isto é, dados que variam em natureza, formato e origem, podendo incluir informações numéricas, categóricas, textuais e até temporais. Exemplos comuns são dados cadastrais, histórico de crédito, renda declarada, comportamento de pagamento, localização geográfica e características socioeconômicas dos clientes. Essa diversidade e complexidade dificultam a análise manual ou o uso de métodos tradicionais para identificar padrões relevantes, tornando fundamental o uso de técnicas estatísticas e computacionais mais avançadas para a previsão do risco de crédito.

Frente a esse desafio, técnicas de modelagem preditiva têm se consolidado como ferramentas essenciais para apoiar a tomada de decisão no setor financeiro. Embora métodos baseados em aprendizado de máquina, como árvores de decisão, *random forests* e redes neurais, tenham ganhado destaque na previsão de risco de crédito, a Regressão Logística permanece amplamente utilizada por sua interpretabilidade, simplicidade de implementação e conformidade com exigências regulatórias, uma vez que permite rastreabilidade dos coeficientes, facilita auditorias e está alinhada às diretrizes de transparência exigidas por órgãos como o Banco Central (Banco Central do Brasil, 2023), ainda que sem menção direta à técnica, e por estudos acadêmicos que destacam sua aceitação em ambientes regulados (Farias, 2023).

Neste trabalho, busca-se explorar o uso da Regressão Logística penalizada pela técnica LASSO *(Least Absolute Shrinkage and Selection Operator)* como ferramenta central na predição de risco de crédito, com atenção especial ao tratamento de dados desbalanceados, uma característica comum nesse tipo de problema. A escolha do modelo se justifica por sua solidez teórica, capacidade de realizar seleção automática de variáveis, e aptidão para estimar probabilidades com base em variáveis socioeconômicas e comportamentais relevantes. Além disso, sua interpretabilidade e rastreabilidade o tornam compatível com exigências regulatórias. A proposta inclui a avaliação do desempenho por métricas apropriadas, como Acurácia Balanceada, AUC (Área Sob a Curva ROC) e F1 Score, com o intuito de garantir resultados robustos e comparáveis.

Ao direcionar a análise exclusivamente para a Regressão Logística, espera-se oferecer uma abordagem concisa, estatisticamente rigorosa e orientada à aplicação prática, contribuindo tanto para o avanço acadêmico quanto para a gestão mais eficiente do risco de crédito por parte das instituições financeiras. Cabe destacar que este estudo se restringe à modelagem do risco de crédito, não abordando outros tipos de risco financeiro, como risco de mercado ou risco de liquidez, que possuem características e metodologias específicas de análise.

1.1 Objetivo Geral

Este trabalho tem como objetivo avaliar a eficácia da regressão logística na modelagem e predição do risco de crédito, utilizando um conjunto de dados de empréstimos com distribuição desbalanceada entre clientes adimplentes e inadimplentes.

1.2 Objetivos Específicos

a) Avaliar o desempenho do modelo por meio de métricas apropriadas para dados desbalanceados (AUC, Acurácia Balanceada e F1 Score);
b) Investigar o impacto do desbalanceamento na performance da regressão logística e a eficácia de estratégias como alteração do *cut-off* (limiar de decisão).

Para atingir esses objetivos será fundamental considerar as particularidades dos conjuntos de dados desbalanceados, típicos em problemas de risco de crédito (Furriel, 2023; Andrade; Silva, 2021). Isso nos leva à importância de estratégias específicas para lidar com essa questão, conforme será discutido na justificativa a seguir.

1.3 Justificativa

Este trabalho se justifica pela relevância do risco de crédito no contexto financeiro, sendo um dos principais desafios enfrentados por instituições financeiras, *fintechs* e investidores. A capacidade de prever com precisão o risco associado a diferentes perfis de clientes permite não apenas mitigar perdas financeiras, mas também viabilizar a concessão de crédito de forma mais segura e inclusiva (Vieira, 2023).

A Regressão Logística é amplamente adotada em estudos de risco de crédito devido à sua robustez estatística, facilidade de implementação e elevada interpretabilidade dos coeficientes estimados. Por ser um modelo amplamente aceito em ambientes regulatórios, ela se destaca como uma ferramenta confiável para a tomada de decisão em instituições financeiras. No entanto, um desafio recorrente nessa área é o desbalanceamento das classes, com uma proporção significativamente menor de inadimplentes em relação aos adimplentes, o que pode comprometer a eficácia do modelo e gerar viés na classificação (Furriel, 2023).

A literatura destaca amplamente essa dificuldade. Estudos como os de He & Garcia (2009), Chawla *et al.* (2002) e Japkowicz & Stephen (2002) enfatizam que métricas tradicionais, como a acurácia simples, podem ser enganosas em cenários desbalanceados, tornando fundamental a utilização de métricas como a AUC, o F1 Score e a acurácia balanceada para uma avaliação mais realista e robusta.

Justifica-se, portanto, a realização deste estudo com foco exclusivo na Regressão Logística, visando não apenas sua aplicação prática à predição do risco de crédito, mas também a

análise crítica do impacto do desbalanceamento de classes e das estratégias possíveis para sua mitigação.

Além disso, este estudo contribui ao explorar a calibragem do *cut-off* como ferramenta para melhorar o desempenho preditivo em contextos de desbalanceamento entre classes. A análise busca destacar quais abordagens oferecem o melhor equilíbrio entre sensibilidade e especificidade. Ao preencher essa lacuna, o trabalho fornece *insights* práticos para a modelagem de crédito em ambientes financeiros complexos.

2 REFERENCIAL TEÓRICO

Este capítulo apresenta os principais conceitos e fundamentos relacionados ao risco de crédito e à sua modelagem estatística. Inicialmente, são discutidas as características do mercado de crédito e a importância da gestão de carteiras para a sustentabilidade das instituições financeiras. Em seguida, abordam-se as classificações das carteiras de crédito e os modelos de rentabilidade utilizados para avaliar o desempenho dessas operações. Por fim, são exploradas as técnicas de modelagem preditiva, com destaque para a regressão logística e os desafios impostos por bases de dados desbalanceadas.

O risco de crédito está diretamente relacionado ao custo de crédito, que representa o preço pago pelo tomador para acessar recursos financeiros. Esse custo inclui não apenas os juros cobrados, mas também encargos, tarifas e o prêmio de risco embutido na operação, ou seja, a compensação exigida pela instituição financeira para assumir o risco de inadimplência. Quanto maior a percepção de risco, maior tende a ser o custo de crédito, impactando diretamente a viabilidade da concessão e a inclusão financeira (StoneX, 2023).

2.1 Risco de Crédito e Estrutura do Mercado

O mercado de crédito tem como principal função suprir as necessidades de recursos financeiros de curto e médio prazos por parte dos diferentes agentes econômicos, tanto pessoas físicas quanto jurídicas. Esse suprimento ocorre por meio da concessão de crédito, empréstimos e financiamentos, geralmente realizados por instituições financeiras bancárias, como bancos comerciais e múltiplos, dentro de um modelo de especialização do Sistema Financeiro Nacional. Além disso, a atuação dos bancos tem se expandido, passando da simples intermediação financeira para uma maior diversificação de produtos e serviços oferecidos ao mercado. Também são consideradas parte do mercado de crédito as operações de financiamento de bens de consumo duráveis, realizadas por sociedades financeiras, ampliando a atuação para instituições não bancárias na oferta de crédito de médio prazo aos consumidores (Assaf Neto, 2018).

A gestão de carteiras de crédito desempenha um papel crucial no desempenho das organizações, pois permite uma alocação eficiente dos recursos disponíveis. Diante da escassez de recursos, tanto empresas quanto indivíduos precisam tomar decisões estratégicas que maximizem os resultados, como a definição do conjunto de produtos a ser comercializado ou a escolha

dos investimentos mais adequados para ampliar o patrimônio (Bordignon, 2020). Diversas técnicas têm sido desenvolvidas para auxiliar nessa gestão, como a Teoria Moderna de Portfólio de Markowitz, o Valor em Risco (*Value at Risk*, VaR), o *CreditMetrics* (modelo desenvolvido pelo J.P. Morgan para mensurar risco de crédito), o KMV (Kealhofer, McQuown and Vasicek, voltado à estimativa da probabilidade de inadimplência), a Perda Esperada e o RAROC (*Risk-Adjusted Return on Capital*, Retorno Ajustado ao Risco sobre o Capital). Essas metodologias oferecem ferramentas para avaliação de desempenho e risco em carteiras de crédito, permitindo uma alocação eficiente de recursos e uma melhor compreensão dos riscos associados às operações (Schuster; Reis, 2024; Jabôr, 2006; Neto, 2011).

As operações integrantes da carteira de crédito de uma instituição financeira podem ser organizadas em distintas categorias, com o objetivo de facilitar sua gestão e monitoramento, levando em conta aspectos como o tipo de produto ou o grau de risco envolvido. Dentre as classificações mais usuais, destacam-se as carteiras de: pessoa física, crédito direto ao consumidor (CDC), cheque especial, empréstimos pessoais, cartões de crédito, pessoa jurídica, desconto de duplicatas, conta garantida, *hot money*, financiamentos de longo prazo e operações de repasse. Essas carteiras podem ser desmembradas ainda mais, como no caso de uma carteira de empréstimos pessoais estruturada por faixas de rating de crédito ou conforme o tipo de garantia associada (Securato, 2015).

O risco de crédito pode ser definido como a possibilidade de uma contraparte não cumprir suas obrigações financeiras conforme o acordado, resultando em perdas para o credor. Segundo Saunders & Allen (2002), esse risco é inerente a qualquer operação de crédito e é considerado uma das principais fontes de instabilidade financeira em instituições bancárias e no mercado financeiro como um todo. A gestão do risco de crédito envolve a identificação, análise e mitigação das possíveis perdas, utilizando ferramentas como análise de crédito, modelos de scoring e técnicas de diversificação de carteira, que visam reduzir a exposição a perdas significativas.

A gestão eficaz do risco de crédito é fundamental para a estabilidade financeira das instituições, pois permite minimizar perdas potenciais decorrentes da inadimplência dos tomadores de empréstimos. De acordo com o estudo de Brito & Neto (2008), a implementação de sistemas de classificação de risco baseados em variáveis contábeis tem se mostrado eficaz na avaliação da solvência das empresas, auxiliando na tomada de decisões sobre concessão de crédito. Esses

sistemas permitem uma análise mais precisa do perfil de risco dos tomadores, contribuindo para a redução da exposição ao risco de crédito.

Nesse contexto, modelos de risco de crédito desempenham um papel crucial ao estimar a probabilidade de inadimplência, um dos principais determinantes na previsão de perdas financeiras. Essas estimativas, geralmente realizadas por meio de técnicas estatísticas, como os modelos de crédito apresentados por Bussmann *et al.* (2021), permitem às instituições financeiras avaliar com maior precisão o risco associado a diferentes tomadores e ajustar suas estratégias de crédito de forma proativa.

2.2 Técnicas de Aprendizado de Máquina

Sarker (2021) apresenta uma revisão abrangente das principais abordagens de aprendizado de máquina, destacando suas aplicações em cenários do mundo real e tendências de pesquisa. A partir dessa base teórica, serão discutidos os métodos relevantes para a modelagem de risco de crédito.

Os algoritmos de aprendizado de máquina são classificados em quatro categorias principais: aprendizado supervisionado, não supervisionado, semi-supervisionado e por reforço. A seguir, são apresentadas essas categorias e suas aplicações.

a) **Aprendizado Supervisionado:** Nesse tipo de aprendizado, o algoritmo é treinado com dados rotulados, ou seja, exemplos em que a resposta correta já é conhecida, para que ele aprenda a associar padrões de entrada às saídas desejadas. Tarefas típicas incluem classificação e regressão, como prever a classe de um item ou estimar valores contínuos com base em variáveis explicativas. No contexto da classificação, o desempenho dos modelos é frequentemente avaliado por meio da matriz de confusão (Tabela 2.1), que resume os acertos e erros do modelo em relação às classes reais. Exemplos incluem a classificação de sentimentos em textos ou a previsão do valor de ações.

b) **Aprendizado Não Supervisionado:** Utilizado para analisar dados não rotulados, esse tipo de aprendizado busca identificar padrões e estruturas nos dados sem a necessidade de intervenção humana. Dentre as principais aplicações estão:

   - **Agrupamento (*clustering*):** técnica que agrupa dados semelhantes entre si, frequentemente utilizada para segmentar clientes com base em padrões de comportamento;

- **Redução de dimensionalidade:** método usado para simplificar conjuntos de dados com muitas variáveis, preservando ao máximo as informações essenciais, como no uso da Análise de Componentes Principais (PCA);
- **Detecção de anomalias:** técnica que identifica padrões fora do comum em conjuntos de dados, sendo útil, por exemplo, para detectar transações suspeitas em sistemas financeiros.

c) **Aprendizado Semi-Supervisionado:** Combina dados rotulados e não rotulados, sendo útil quando há escassez de dados rotulados. Essa abordagem é eficaz para melhorar os resultados preditivos em situações com poucos dados rotulados e muitos dados não rotulados. Exemplos de aplicação incluem a tradução de idiomas e a detecção de fraudes.

d) **Aprendizado por Reforço:** Nesse tipo de aprendizado, um agente aprende a tomar decisões com base em recompensas ou penalidades, buscando maximizar a recompensa ao longo do tempo. Essa abordagem é especialmente eficaz em ambientes dinâmicos e sequenciais, como robótica, direção autônoma e logística. No entanto, sua aplicação tende a ser mais vantajosa em problemas complexos e com múltiplas etapas de decisão, sendo geralmente menos recomendada para tarefas diretas ou com baixa variabilidade de estados.

Os algoritmos de aprendizado de máquina utilizam dados para identificar padrões relacionados a indivíduos, processos de negócios, transações, eventos, entre outros. Esses dados podem ser classificados em diferentes tipos, os quais influenciam diretamente a escolha das técnicas de aprendizado de máquina.

2.3 Matriz de confusão

Os resultados de um modelo de classificação binária podem ser avaliados com base em quatro categorias principais: verdadeiro positivo (TP), quando a instância positiva é corretamente classificada como tal; falso positivo (FP), quando uma instância negativa é incorretamente classificada como positiva; verdadeiro negativo (TN), quando uma instância negativa é corretamente classificada; e falso negativo (FN), quando uma instância positiva é classificada incorretamente como negativa.

Essas quatro categorias podem ser organizadas em uma matriz de confusão, que é uma ferramenta utilizada para visualizar o desempenho de um modelo de classificação, comparando as previsões do modelo com os valores reais observados, mostrada na tabela 2.1

Tabela 2.1 – Matriz de Confusão

	Classe Predita: Positiva	**Classe Predita: Negativa**
Classe Real: Positiva	Verdadeiro Positivo (TP)	Falso Negativo (FN)
Classe Real: Negativa	Falso Positivo (FP)	Verdadeiro Negativo (TN)

Fonte: Elaborado pelo autor.

2.3.1 Dados desbalanceados

Em problemas de classificação binária, como a previsão de inadimplência, é comum que as classes estejam desigualmente representadas. Um conjunto de dados é considerado desbalanceado quando a proporção entre as classes é significativamente desigual, o que pode comprometer a capacidade do modelo de identificar corretamente a classe minoritária (He; Garcia, 2009; Japkowicz; Stephen, 2002). Desequilíbrios desse tipo são frequentes em aplicações do mundo real, especialmente em domínios como detecção de fraudes, diagnóstico médico e risco de crédito.

Por exemplo, considere um conjunto de dados com informações sobre o comportamento financeiro de 100 indivíduos, dos quais 95 são adimplentes e apenas 5 são inadimplentes. Nesse cenário, a classe majoritária é representada pelos adimplentes, enquanto a classe minoritária é composta pelos inadimplentes. Se utilizarmos um modelo de aprendizado de máquina que simplesmente classifique todos os indivíduos como adimplentes, ele obterá uma acurácia preditiva de 95%, uma vez que 95 de 100 classificações estarão corretas. No entanto, esse modelo falha completamente em identificar qualquer caso de inadimplência, o que torna sua utilidade limitada em aplicações práticas. Essa situação demonstra que a acurácia preditiva não é uma métrica apropriada em cenários de dados desbalanceados, especialmente quando a identificação correta da classe minoritária é de grande importância (He; Garcia, 2009).

Métricas como a curva de *Receiver Operating Characteristic* (ROC), a área sob a curva (AUC) e a acurácia balanceada são amplamente utilizadas para avaliar o desempenho de classificadores em situações de desequilíbrio. Essas ferramentas permitem um melhor entendimento

do compromisso entre taxas de verdadeiros positivos e falsos positivos, auxiliando na seleção de modelos que sejam eficazes para ambas as classes. Técnicas como a superamostragem da classe minoritária e a subamostragem da classe majoritária têm se mostrado eficazes para melhorar o desempenho de modelos em conjuntos de dados desbalanceados (Chawla *et al.*, 2002).

2.4 Tipos de dados

A disponibilidade e a organização dos dados são cruciais para a construção de modelos de aprendizado de máquina. De modo geral, os dados podem ser classificados em três categorias principais, conforme sua estrutura e organização (Sarker, 2021):

a) **Dados estruturados:** São dados organizados segundo um esquema fixo e bem definido, geralmente armazenados em tabelas com linhas e colunas, no qual cada coluna representa um atributo específico e cada linha corresponde a uma instância do conjunto de dados. Essa estrutura permite que as informações sejam facilmente acessadas, processadas e analisadas por meio de consultas em bancos de dados relacionais.

b) **Dados não estruturados:** São dados que não seguem um esquema predefinido ou uma organização rígida, dificultando sua indexação e análise automatizada. Eles podem apresentar diferentes formatos e não possuem uma estrutura fixa que permita uma categorização direta. Como resultado, a extração e o processamento dessas informações geralmente exigem técnicas avançadas, como processamento de linguagem natural e visão computacional.

c) **Dados semi-estruturados:** Situam-se entre os dados estruturados e os não estruturados, pois contêm elementos organizacionais que permitem certo nível de categorização, mas não seguem um modelo tabular rígido. Eles frequentemente apresentam marcações ou hierarquias que facilitam a identificação e extração de informações sem a necessidade de um banco de dados relacional tradicional.

d) **Metadados:** São dados que descrevem outros dados, fornecendo contexto adicional para sua interpretação e organização. Eles podem indicar propriedades como fonte, autor, data de criação ou formato, desempenhando um papel essencial na categorização e recuperação eficiente das informações.

A disponibilidade e a organização dos dados são cruciais para a construção de modelos de aprendizado de máquina. No contexto do risco de crédito, os dados utilizados geralmente são

estruturados e contêm informações como histórico de pagamentos, renda, score de crédito, nível de endividamento, número de empréstimos, entre outros atributos financeiros e demográficos dos clientes.

Além dos dados estruturados, também podem ser incorporadas fontes de dados com algum grau de organização interna, mas que não seguem esquemas tabulares rígidos. Esses dados ampliam a capacidade do modelo de capturar nuances comportamentais e operacionais dos clientes. Em menor escala, informações não estruturadas, como descrições textuais de análises de crédito ou comunicações com clientes, podem fornecer contexto adicional, enriquecendo a análise e contribuindo para uma compreensão mais abrangente do perfil de risco.

Dessa forma, a diversidade dos dados disponíveis no setor financeiro reforça a importância da sua organização e pré-processamento adequado para uma modelagem eficiente do risco de crédito.

2.4.1 Aumentação de Dados (*Data Augmentation*)

As técnicas de aumentação de dados são amplamente utilizadas para lidar com conjuntos de dados desbalanceados, visando aumentar a representatividade das classes minoritárias e reduzir o viés em favor da classe majoritária. Esse processo busca tornar os modelos mais equilibrados e precisos. Algumas dessas técnicas incluem a sobreamostragem, a subamostragem e o SMOTE *(Synthetic Minority Over-sampling Technique)*, que têm se mostrado eficazes em diferentes contextos. Diversos estudos demonstram os benefícios dessas abordagens na melhoria da acurácia e do balanceamento da performance dos classificadores (Fayz, Rizka & Maghraby (2018); Tan *et al.* (2019); Rupapara *et al.* (2021)).

No entanto, estudo recente de Assunção, Izbicki & Prates (2024) mostra que tais técnicas podem não ser essenciais, e que resultados equivalentes, ou até superiores, podem ser obtidos apenas com a escolha apropriada do limiar de classificação, sem a necessidade de aumentação de dados.

A sobreamostragem é frequentemente aplicada porque a maioria dos algoritmos de aprendizado de máquina, por padrão, classifica novas instâncias com base na maior probabilidade estimada. No entanto, essa estratégia pode ser subótima. Assunção, Izbicki & Prates (2024) demonstraram que o simples ajuste do limiar de decisão em classificadores treinados exclusivamente com os dados originais pode resultar em um desempenho comparável, ou até

superior, ao obtido com técnicas de aumentação. Isso evidencia que a sobreamostragem nem sempre proporciona as vantagens que lhe são comumente atribuídas.

Adicionalmente, a escolha de um limiar otimizado com base nas métricas de interesse, aplicada diretamente ao conjunto de dados desbalanceado, mostra-se suficiente para assegurar uma boa precisão preditiva, dispensando a necessidade de técnicas de aumentação ((Assunção; Izbicki; Prates, 2024)).

2.4.2 Subamostragem (*Undersampling*)

A técnica de subamostragem consiste em reduzir a quantidade de instâncias da classe majoritária, buscando equilibrar a distribuição das classes no conjunto de dados. Essa abordagem é simples e eficaz, especialmente quando o número de observações da classe majoritária é muito superior ao da classe minoritária. A principal desvantagem da subamostragem é a possível perda de informações relevantes, uma vez que instâncias importantes podem ser removidas do conjunto de treinamento (Dubey *et al.*, 2014). Isso pode impactar negativamente a capacidade preditiva do modelo, principalmente em cenários com grande quantidade de dados ou alta variabilidade na classe majoritária.

2.4.3 Sobreamostragem (*Oversampling*)

A sobreamostragem, por sua vez, busca aumentar a representatividade da classe minoritária. A forma mais simples de realizar essa técnica é replicando exemplos já existentes dessa classe. No entanto, essa estratégia pode levar ao sobreajuste (*overfitting*), pois o modelo pode memorizar as instâncias duplicadas, comprometendo sua capacidade de generalização. Por esse motivo, métodos mais sofisticados foram desenvolvidos para evitar essa limitação, como o SMOTE (Chawla *et al.*, 2002).

2.5 SMOTE (Synthetic Minority Over-sampling Technique)

Em 2002, Chawla *et al.* (2002) propuseram uma abordagem para lidar com conjuntos de dados desbalanceados, buscando evitar o sobreajuste causado pela simples replicação de instâncias minoritárias. Essa técnica, denominada SMOTE (*Synthetic Minority Over-sampling*

Technique), baseia-se na geração de novas amostras sintéticas em vez de apenas duplicar exemplos existentes.

O SMOTE cria novos exemplos ao interpolar entre instâncias reais da classe minoritária, utilizando seus k vizinhos mais próximos. Esses vizinhos são determinados com base na distância euclidiana no espaço das características, garantindo que as amostras sintéticas estejam distribuídas de maneira mais homogênea. Dessa forma, o método expande a região de decisão da classe minoritária, tornando-a mais generalizada (Chawla *et al.*, 2002).

É importante destacar que ao gerar novas instâncias sintéticas, o SMOTE modifica diretamente o conjunto de dados original. Embora essa expansão possa ajudar no balanceamento das classes, a criação de dados artificiais não necessariamente resulta em uma melhoria real da capacidade preditiva do modelo. Em algumas situações, isso pode introduzir ruído ou padrões artificiais que não refletem fielmente a distribuição dos dados reais, impactando negativamente a performance ou interpretabilidade do modelo.

Desde seu lançamento, o SMOTE tornou-se uma técnica fundamental na comunidade de aprendizado de máquina, levando ao desenvolvimento de diversas extensões e abordagens híbridas. Algumas pesquisas combinam o SMOTE com subamostragem da classe majoritária para lidar com desbalanceamentos extremos (Fernández *et al.*, 2018). Seu impacto se estende a diversas aplicações, como diagnóstico médico, reconhecimento facial, redes sociais e engenharia de software, tornando-se uma referência no pré-processamento de dados desbalanceados.

Apesar de sua ampla utilização, o algoritmo SMOTE ainda carece de uma fundamentação teórica completamente consolidada (Xu *et al.*, 2020).

2.6 Validação cruzada e ajuste de hiperparâmetros

A validação cruzada é uma técnica utilizada para avaliar a performance de modelos de aprendizado de máquina e melhorar sua generalização. O método mais comum é o k-fold, no qual o conjunto de dados é dividido em k partições iguais. O modelo é treinado k vezes, cada vez usando $k-1$ partições para treinamento e a partição restante para teste. Esse processo é repetido para todas as partições, e a média dos resultados obtidos é calculada, proporcionando uma avaliação mais confiável e reduzindo o risco de sobreajuste *(overfitting)*. A validação cruzada ajuda a garantir que o modelo tenha um desempenho robusto em dados não vistos durante o treinamento (Ramezan; Warner; Maxwell, 2019).

Além do k-fold, existem outras variações, como a validação *leave-one-out*, na qual cada amostra é utilizada uma vez como conjunto de teste, o que pode ser mais preciso, porém mais lento. A validação Monte Carlo, por sua vez, realiza divisões aleatórias dos dados, permitindo que algumas amostras sejam usadas múltiplas vezes, enquanto outras não sejam utilizadas. Esses métodos são úteis para testar o modelo de forma mais abrangente, proporcionando uma melhor estimativa de seu desempenho em dados reais e aumentando sua capacidade de generalização (Ramezan; Warner; Maxwell, 2019).

O ajuste de hiperparâmetros é outro componente fundamental na construção de modelos preditivos eficazes. Hiperparâmetros são configurações externas ao modelo que influenciam seu comportamento, como o valor de penalização na regressão logística com regularização LASSO, o número de vizinhos em algoritmos *k-nearest neighbors*, ou a profundidade máxima em árvores de decisão. A escolha adequada desses valores pode impactar diretamente o desempenho do modelo. Técnicas como busca em grade (*grid search*) e busca aleatória (*random search*) são comumente utilizadas em conjunto com validação cruzada para identificar a combinação de hiperparâmetros que maximiza o desempenho preditivo, evitando tanto o sobreajuste quanto a subajuste (Bergstra; Bengio, 2012).

2.6.1 *Grid Search*

O grid search (GS) é uma técnica simples e eficiente para otimização de hiperparâmetros em modelos de aprendizado de máquina. Nessa abordagem, todas as combinações de hiperparâmetros geradas a partir de um conjunto finito de valores definidos pelo usuário são testadas em uma grade. O GS treina o modelo para cada combinação, utilizando o produto cartesiano desses valores. O desempenho de cada configuração é avaliado com base em um conjunto de validação separado do conjunto de treinamento. Em seguida, o algoritmo identifica os hiperparâmetros que apresentam o melhor desempenho durante a validação, adotando-os como os valores finais do modelo (Fuadah; Pramudito; Lim, 2023).

2.7 Métricas de Avaliação

Nesta seção, são apresentadas as métricas utilizadas para avaliar o desempenho dos modelos de aprendizado de máquina aplicados à modelagem do risco de crédito. As métricas

selecionadas são: Acurácia, Acurácia Balanceada, Precisão, Recall, F1-Score, AUC (Área Sob a Curva ROC) e Kappa, devido à sua relevância na análise de classificadores, especialmente em cenários com dados desbalanceados.

2.7.1 Acurácia

A acurácia é uma métrica que avalia o quão preciso um modelo é em relação aos dados em geral (Coelho; Amorim; Camargos, 2021). Ela é obtida pela divisão do número de unidades corretamente classificadas, verdadeiros positivos (TP) e verdadeiros negativos (TN), pelo número total de previsões feitas pelo classificador, levando em consideração também os falsos positivos (FP) e falsos negativos (FN). O valor da acurácia varia entre 0 e 1, sendo que quanto mais próximo de 1, melhor o desempenho do modelo.

$$\text{ACC} = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.1}$$

2.7.2 Acurácia Balanceada

A acurácia balanceada considera o desempenho do modelo em cada classe individualmente e calcula a média entre elas. Isso é particularmente importante em situações de desequilíbrio de classes, como neste estudo, onde o número de adimplentes é significativamente maior que o de inadimplentes. Em casos de desbalanceamento de dados, a acurácia simples pode ser enganosa, já que o modelo tende a favorecer a classe majoritária (adimplentes), resultando em uma avaliação distorcida do desempenho. A acurácia balanceada, por outro lado, oferece uma medida mais equitativa, avaliando a performance em ambas as classes de forma justa e proporcionando uma visão mais completa da capacidade preditiva dos modelos utilizados (Brodersen *et al.*, 2010). A acurácia balanceada é muito importante para dados desbalanceados.

A fórmula da acurácia balanceada é dada por:

$$\text{Acurácia Balanceada} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{2.2}$$

2.7.3 Curva ROC

As curvas ROC (*Receiver Operating Characteristic*) podem ser interpretadas como representações da família de melhores limiares de decisão para custos relativos de verdadeiros positivos (TP) e falsos positivos (FP). Em uma curva ROC:

a) O eixo X representa o percentual de falsos positivos (FP), dado por:

$$FP = \frac{\text{FP}}{\text{TN} + \text{FP}} \tag{2.3}$$

b) O eixo Y representa o percentual de verdadeiros positivos (TP), dado por:

$$TP = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.4}$$

O ponto ideal em uma curva ROC seria $(0, 100)$, em que todos os exemplos positivos são classificados corretamente e nenhum exemplo negativo é classificado incorretamente como positivo. A linha $y = x$ representa o cenário de adivinhação aleatória das classes. Uma maneira de construir a curva ROC é manipulando o equilíbrio das amostras de treinamento para cada classe no conjunto de treinamento.

A Área sob a curva ROC (AUC) é uma métrica amplamente utilizada para avaliar o desempenho de classificadores. Ela é útil porque é independente do critério de decisão selecionado e das probabilidades a priori. A AUC permite comparar classificadores e estabelecer uma relação de dominância entre eles (Chawla *et al.*, 2002). O valor da AUC reflete a capacidade do modelo em discriminar entre classes, sendo interpretado da seguinte forma:

a) Valores menores ou iguais a 0,5 indicam que o modelo não tem capacidade de discriminação;
b) Valores entre 0,5 e 0,7 indicam discriminação baixa;
c) Valores entre 0,7 e 0,8 indicam discriminação aceitável;
d) Valores entre 0,8 e 0,9 indicam discriminação excelente; e
e) Valores acima de 0,9 indicam discriminação excepcional (FÁVERO *et al.*, 2009).

A Figura 2.1 apresenta uma representação gráfica típica da curva ROC, destacando a relação entre a taxa de verdadeiros positivos e a taxa de falsos positivos. A linha diagonal indica

o desempenho esperado em um cenário de classificação aleatória, enquanto a curva acima dela ilustra um modelo com maior capacidade discriminativa.

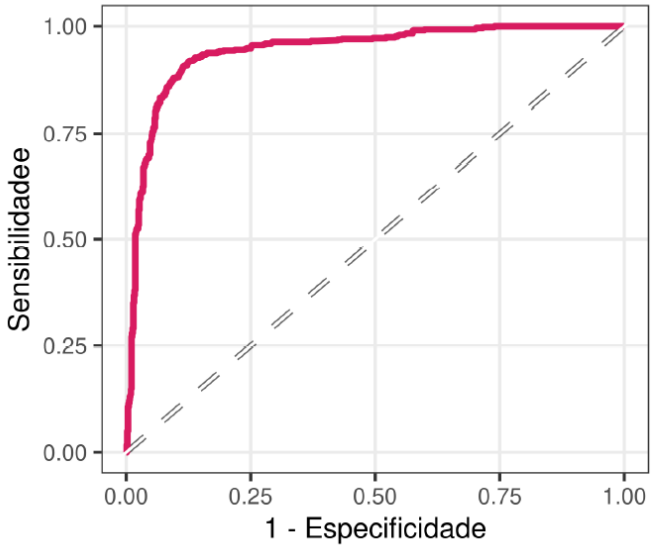

Figura 2.1 – Curva ROC ilustrativa. Fonte: (Izbicki; Santos, 2020).

2.7.4 Precisão, Recall e F1 Score

Precisão, Recall e F1 Score são amplamente utilizadas para avaliar o desempenho de modelos de classificação binária, especialmente em cenários com dados desbalanceados (Miller *et al.*, 2024).

$$\text{Precisão} = \frac{TP}{TP+FP}, \quad \text{Recall} \quad = \frac{TP}{TP+FN}, \quad F1 = \frac{2 \times P \times R}{P+R} \tag{2.5}$$

onde TP, FP e FN correspondem às categorias da matriz de confusão (Seção 2.3), e P e R representam, respectivamente, a precisão e o recall calculados.

A **Precisão** (ou valor preditivo positivo) reflete a proporção de amostras previstas como "positivas"que realmente pertencem à classe positiva. Um valor de 100% de precisão indica que não houve amostras incorretamente classificadas como positivas. Por outro lado, o **Recall** (ou sensibilidade) mede a proporção de amostras positivas que foram corretamente identificadas pelo modelo. Um recall de 100% indica que nenhuma amostra positiva foi erroneamente classificada como negativa.

O **F1 Score** representa a média harmônica entre precisão e recall, sendo uma métrica que exige um equilíbrio elevado entre ambos os valores para ser alto. A relevância de precisão e recall varia de acordo com o contexto do problema. Por exemplo, em diagnósticos médicos, classificar erroneamente um caso positivo como negativo pode ser mais prejudicial, o que torna o recall mais relevante. Por outro lado, em estudos genômicos focados em identificar variantes genéticas, garantir que a maioria das variantes identificadas sejam corretas pode ser mais importante, priorizando assim a precisão.

2.7.5 Estatística Kappa

As coeficiente Kappa de Cohen (κ) é uma medida que avalia quantas instâncias foram corretamente classificadas por um modelo de aprendizado de máquina em comparação com os dados considerados como verdade básica, ajustando para a precisão de um classificador aleatório como ponto de referência (Vujović *et al.*, 2021).

A precisão de um classificador aleatório é dada por $1/k$, em que k é o número de classes no conjunto de dados. No caso de uma classificação binária ($k = 2$), a precisão esperada é de 50%.

O coeficiente Kappa é calculado pela seguinte fórmula:

$$K = \frac{p_0 - p_e}{1 - p_e} \tag{2.6}$$

Onde:

a) p_0: precisão total do modelo (proporção de casos corretamente classificados).
b) p_e: precisão esperada (precisão que seria obtida por um classificador aleatório).

No caso de classificação binária, p_e é a soma das probabilidades de concordância esperada para cada classe:

$$p_e = p_{e1} + p_{e2}$$

E as probabilidades p_{e1} e p_{e2} são calculadas como:

$$p_{e1} = \text{proporção da classe 1} \times \text{proporção prevista para a classe 1}$$

$$p_{e2} = \text{proporção da classe 2} \times \text{proporção prevista para a classe 2}$$

O coeficiente Kappa varia de -1 a 1, onde:

a) $\kappa = 1$: Concordância perfeita.

b) $\kappa = 0$: Concordância equivalente ao acaso.

c) $\kappa < 0$: Concordância inferior ao esperado pelo acaso.

As métricas de avaliação são fundamentais para analisar o desempenho de modelos de classificação, especialmente em contextos de dados desbalanceados. O Quadro 2.1 apresenta as principais métricas utilizadas neste trabalho, bem como suas respectivas fórmulas e descrições.

Quadro 2.1 – Métricas de avaliação e suas fórmulas

Métrica	**Fórmula**	**Descrição**
Acurácia	$\frac{TP+TN}{TP+TN+FP+FN}$	Proporção de previsões corretas (positivas e negativas).
Precisão	$\frac{TP}{TP+FP}$	Entre os positivos previstos, quantos são realmente positivos.
Revocação (Recall)	$\frac{TP}{TP+FN}$	Entre os positivos reais, quantos foram corretamente identificados.
F1 Score	$2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}}$	Média harmônica entre precisão e revocação.
Especificidade	$\frac{TN}{TN+FP}$	Capacidade de identificar corretamente os negativos.
AUC-ROC	—	Área sob a curva ROC (não possui fórmula fechada).

2.7.6 Risco Relativo

Um valor $\pi_1 - \pi_2$ de tamanho fixo pode ter maior importância quando ambos os π_i estão próximos de 0 ou 1 do que quando não estão. Em um estudo comparando dois tratamentos sobre a proporção de sujeitos que morrem, a diferença entre 0,010 e 0,001 é mais relevante do que a diferença entre 0,410 e 0,401, embora ambas sejam 0,009. Nesses casos, a razão de proporções também é informativa.

O risco relativo é definido como a razão entre duas probabilidades, conforme a equação:

$$\text{Risco Relativo} = \frac{\pi_1}{\pi_2} \tag{2.7}$$

Ele pode ser qualquer número real não negativo. Um risco relativo de 1,0 corresponde à independência. Para as proporções mencionadas, os riscos relativos são $0,010/0,001 = 10,0$ e $0,410/0,401 = 1,02$. Comparar as linhas na segunda categoria de resposta resulta em um risco relativo diferente, $\frac{(1-\pi_1)}{(1-\pi_2)}$.

2.8 Modelos de Classificação

Modelos de classificação são amplamente utilizados em problemas de previsão binária, como a identificação de inadimplência no mercado de crédito. Esses modelos têm como objetivo estimar a probabilidade de ocorrência de um evento com base em um conjunto de variáveis explicativas. No contexto deste estudo, o foco está na regressão logística penalizada, uma técnica estatística que combina interpretabilidade com capacidade preditiva, especialmente útil em cenários com desbalanceamento entre classes. A seguir, são apresentados os fundamentos teóricos e práticos da regressão logística, bem como os métodos de seleção de variáveis e os aspectos diagnósticos do modelo.

2.8.1 Regressão Logística

Considere um conjunto de p variáveis independentes representadas pelo vetor $x = (x_1, x_2, \ldots, x_p)$. Seja $Pr(Y = 1|x) = \pi(x)$ a probabilidade de que o resultado ocorra, dado o valor das variáveis x. O modelo de regressão logística múltipla é dado por (Hosmer; Lemeshow; Sturdivant, 2013)

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}},$$

em que

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

O *logit* do modelo de regressão logística múltipla é dado pela equação:

$$
\begin{aligned}
\text{logit}(\pi(x)) &= \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \log(\pi(x)) - \log(1-\pi(x)) \\
&= \log\left(\frac{e^{g(x)}}{1+e^{g(x)}}\right) - \log\left(1 - \frac{e^{g(x)}}{1+e^{g(x)}}\right) \\
&= \log\left(\frac{e^{g(x)}}{1+e^{g(x)}}\right) - \log\left(\frac{1+e^{g(x)}-e^{g(x)}}{1+e^{g(x)}}\right) \\
&= \log\left(\frac{e^{g(x)}}{1+e^{g(x)}}\right) - \log\left(\frac{1}{1+e^{g(x)}}\right) \\
&= \log\left(\frac{e^{g(x)}}{1+e^{g(x)}} \cdot \frac{1+e^{g(x)}}{1}\right) \\
&= \log(e^{g(x)}) \\
&= g(x)
\end{aligned}
$$

A regressão logística é uma técnica utilizada para modelar a probabilidade de ocorrência de um evento, que pode ser binário (como no caso da inadimplência de crédito) ou multiclasse. Ao invés de prever diretamente a resposta (inadimplência sim ou não), ela estima a probabilidade de ocorrência de cada categoria, utilizando variáveis explicativas. Para o risco de crédito, a probabilidade de inadimplência, dada um conjunto de variáveis como idade, renda, e valor do empréstimo, é modelada pela equação:

$$\text{Pr}(\text{inadimplência} = \text{Sim} \mid \text{variáveis}) = p(\text{variáveis}). \tag{2.8}$$

Sob condições regulares, os estimadores obtidos por máxima verossimilhança são consistentes, assintoticamente normais e eficientes. Isso significa que, para grandes amostras, os coeficientes estimados seguem uma distribuição normal centrada nos verdadeiros valores, com variância dada pela inversa da matriz de informação de Fisher.

2.8.1.1 Importância do cut-off

Embora o valor padrão de corte seja 0,5, essa escolha nem sempre é ideal, especialmente em cenários desbalanceados. A escolha de um ponto de corte mais adequado pode melhorar significativamente métricas como sensibilidade, especificidade e F1 Score. Segundo Encord

(2023), o F1 Score é uma métrica especialmente útil em contextos de desbalanceamento, pois equilibra precisão e recall de forma harmônica, sendo mais informativa que a acurácia em muitos casos. A varredura de *cut-offs* permite identificar o ponto que maximiza a acurácia balanceada (James *et al.*, 2013).

2.8.1.2 Interpretação dos coeficientes

Os coeficientes estimados no modelo logístico representam o impacto de cada variável explicativa sobre o logaritmo da razão de chances (*logit*). Para facilitar a interpretação, é comum aplicar a exponenciação dos coeficientes, obtendo o *odds ratio*, dado por e^{β_i}. Esse valor indica o quanto a chance de inadimplência se multiplica para cada unidade adicional da variável x_i, mantendo as demais constantes. Por exemplo, se $e^{\beta_1} = 1{,}5$, isso significa que um aumento unitário em x_1 está associado a um aumento de 50% na chance de inadimplência.

Esses métodos são particularmente úteis em estudos de crédito, pois permitem identificar quais características de clientes são mais determinantes para prever a inadimplência, como demonstrado em um estudo sobre a previsão de inadimplência baseado em variáveis como renda e histórico de crédito (James *et al.*, 2013).

2.8.1.3 Estimação dos parâmetros

Os coeficientes do modelo logístico são estimados por meio do método da máxima verossimilhança, que busca os valores que tornam os dados observados mais prováveis sob o modelo. Esse processo é realizado por algoritmos de otimização numérica, amplamente implementados em softwares como R e Python. Embora o cálculo envolva derivadas e iterações, seu uso é transparente para o usuário final e garante estimativas consistentes e eficientes.

A função de log-verossimilhança associada ao modelo é dada por

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i) \right]. \tag{2.9}$$

Como não existe solução em forma fechada para os estimadores, $\ell(\beta)$ é maximizada numericamente, em geral por meio do algoritmo de Newton–Raphson, obtendo-se assim os estimadores de máxima verossimilhança de β, em que $\pi_i = \frac{e^{x_i^\top \beta}}{1+e^{x_i^\top \beta}}$ representa a probabilidade

estimada para a i-ésima observação. Essa função é maximizada numericamente para obter os estimadores dos coeficientes.

A variância dos estimadores pode ser obtida a partir da matriz de informação de Fisher, definida como:

$$\mathscr{I}(\beta) = \sum_{i=1}^{n} \pi_i(1-\pi_i)x_i x_i^\top.$$

Essa matriz é utilizada para construir intervalos de confiança assintóticos e realizar testes de hipótese sobre os coeficientes estimados.

Em prática, muitas implementações utilizam o método *Iteratively Reweighted Least Squares* (IRLS), que otimiza a log-verossimilhança com maior estabilidade numérica.

2.8.2 Modelo Logístico

A função logística, utilizada para modelar a probabilidade $p(X)$ de um evento, garante que a previsão esteja entre 0 e 1 para todos os valores de X. Ela é definida por:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}, \tag{2.10}$$

em que β_0 e β_1 são os coeficientes estimados a partir dos dados de treinamento, e a equação gera uma curva sigmoide que mapeia os valores das variáveis explicativas para uma probabilidade de inadimplência. Essa formulação é amplamente adotada, pois garante que as probabilidades estimadas estejam sempre no intervalo $[0,1]$.

A Figura 2.2 ilustra o comportamento da função logística, que gera uma curva sigmoide ao mapear as variáveis explicativas para probabilidades no intervalo [0,1].

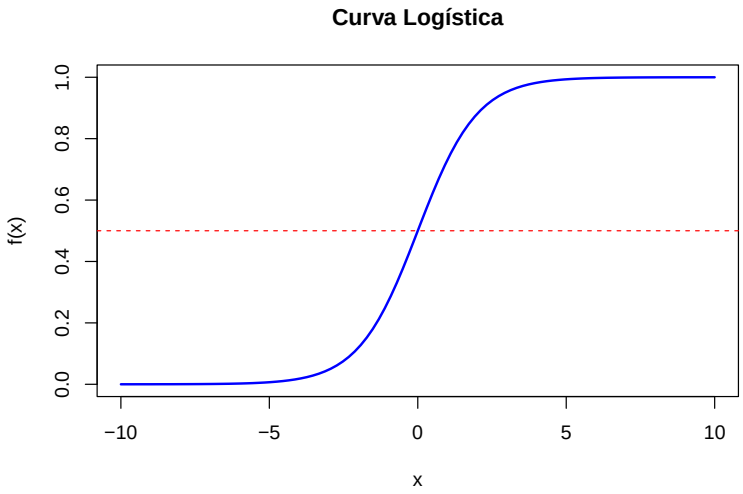

Figura 2.2 – Comportamento da função logística em modelos de classificação binária

Fonte: Do autor (2025).

2.8.3 Seleção de variáveis com *Stepwise*

Uma abordagem adicional utilizada neste estudo para aprimorar o desempenho do modelo foi a aplicação do método de seleção de variáveis *stepwise*. Este procedimento automatiza a inclusão ou exclusão de variáveis com base em critérios estatísticos, como valor-p ou AIC, buscando manter apenas as variáveis mais relevantes para a previsão da inadimplência.

a) **Stepwise Forward:** Começa com um modelo vazio e adiciona variáveis uma a uma, com base em critérios como valor-p ou AIC, até que não haja mais melhorias.
b) **Stepwise Backward:** Começa com todas as variáveis e remove aquelas que mais prejudicam o modelo, com base nos mesmos critérios.

2.8.4 Seleção de Variáveis via Penalização: LASSO

Além dos métodos tradicionais de seleção de variáveis, como o *stepwise*, destaca-se a técnica LASSO (*Least Absolute Shrinkage and Selection Operator*), proposta por Tibshirani (1996). O LASSO realiza simultaneamente estimação e seleção de variáveis ao adicionar uma penalização à soma dos valores absolutos dos coeficientes na função de verossimilhança. Essa penalização, conhecida como regularização L1, tende a reduzir alguns coeficientes a zero, pro-

movendo modelos mais parcimoniosos e interpretáveis, especialmente úteis em contextos com muitas variáveis explicativas ou alta colinearidade.

No contexto da regressão logística, o LASSO é aplicado sobre a função de verossimilhança penalizada, e o parâmetro de regularização λ controla a intensidade da penalização. A escolha adequada de λ é geralmente feita por validação cruzada, utilizando métricas como AUC ou erro de classificação.

Formalmente, o modelo penalizado via LASSO busca maximizar a função:

$$\ell_{\text{pen}}(\beta) = \ell(\beta) - \lambda \sum_{j=1}^{p} |\beta_j|,$$

em que $\ell(\beta)$ é a log-verossimilhança da regressão logística. Valores maiores de λ induzem maior parcimônia, reduzindo mais coeficientes a zero.

Importa destacar que o termo $\sum_{j=1}^{p} |\beta_j|$ não inclui o intercepto β_0, preservando o nível de base do modelo (Hastie, Tibshirani & Friedman (2009a)).

Essa abordagem tem se mostrado eficaz em diversos estudos aplicados, inclusive em cenários de alta dimensionalidade, onde os métodos clássicos enfrentam limitações.

2.8.5 Validação Cruzada para Seleção do Parâmetro λ

A escolha do valor ideal de λ é crítica para o desempenho do modelo, pois um λ muito pequeno não penaliza os coeficientes, enquanto um λ muito grande pode forçar coeficientes importantes a se tornarem zero. O método padrão para determinar o λ ótimo é a validação cruzada.

O processo de validação cruzada para o LASSO geralmente segue os seguintes passos:

1. Divide-se o conjunto de dados de treinamento em k subconjuntos (*folds*) de tamanho aproximadamente igual.

2. Para cada valor de λ em uma grade pré-definida, o modelo é treinado k vezes. Em cada iteração, um *fold* é mantido como conjunto de validação e o modelo é ajustado aos $k-1$ *folds* restantes.

3. A performance do modelo (por exemplo, a acurácia, o AUC ou a acurácia balanceada) é calculada no *fold* de validação.

4. Ao final das k iterações, a média da performance é calculada para cada valor de λ.

5. O valor de λ que resulta na melhor performance média (ou o mais parcimonioso dentro de um desvio padrão do valor máximo) é selecionado como o λ ideal.

Essa abordagem garante que a escolha de λ seja robusta e não superestime o desempenho do modelo, pois o parâmetro é ajustado com base em dados que não foram usados para o treinamento inicial. A validação cruzada para o LASSO tem se mostrado eficaz em diversos estudos aplicados, inclusive em cenários de alta dimensionalidade, onde os métodos clássicos enfrentam limitações.

A escolha de λ também pode levar em conta funções de custo específicas (por exemplo, custo de falsos positivos vs. falsos negativos), ajustando a penalização ao perfil do problema.

2.8.6 Aspectos diagnósticos do modelo

A regressão logística assume que há uma relação linear entre os preditores e o *logit* da variável resposta. Quando essa suposição é violada, o modelo pode apresentar baixa capacidade preditiva. Além disso, a presença de multicolinearidade — ou seja, correlação elevada entre variáveis explicativas — pode inflar as variâncias dos coeficientes, dificultando a interpretação e a estabilidade do modelo.

Outliers também podem influenciar fortemente os coeficientes estimados, especialmente em bases com poucos inadimplentes. Por isso, é recomendável realizar análises de influência, como o uso de gráficos de resíduos ou medidas como o *Cook's Distance*, para identificar observações que distorcem o ajuste.

2.9 Comparação entre modelos de aprendizado de máquina

Antes de analisar o desempenho dos modelos aplicados ao problema de classificação de inadimplência, é importante entender as características e limitações de diferentes técnicas de aprendizado de máquina. Cada método possui vantagens específicas que podem ser mais ou menos adequadas dependendo do tipo de dados e do problema a ser resolvido. A seguir, é apresentada uma comparação entre alguns dos métodos mais comuns de aprendizado de máquina.

Tabela 2.2 – Comparação entre Regressão Logística, Árvores de Decisão, Bagging, Floresta Aleatória e Boosting

Método	Vantagens (Características)	Limitações
Regressão Logística	- Simples e fácil de interpretar. - Adequada para classificação binária. - Probabilidades bem definidas.	- Assume linearidade entre as variáveis independentes e a variável resposta. - Pode ter desempenho ruim em problemas complexos.
Árvores de Decisão	- Fácil de interpretar e visualizar. - Captura relações não lineares. - Rápida para treinar.	- Propensa a sobreajuste em árvores profundas. - Instável (pequenas mudanças nos dados resultam em grandes alterações na árvore).
Bagging	- Reduz o sobreajuste combinando vários modelos. - Estável devido à agregação. - Independência dos modelos paralelos.	- Perda de interpretabilidade. - Depende da qualidade das árvores individuais.
Floresta Aleatória	- Melhora significativamente o desempenho das árvores individuais. - Menos propensa a sobreajuste devido à média das previsões. - Reduz a variância.	- Perda de interpretabilidade em relação a uma única árvore. - Pode ser computacionalmente intensiva com muitas árvores.
Boosting	- Alta performance, especialmente em dados tabulares. - Corrige erros iterativamente. - Modelos como XGBoost são eficientes e poderosos.	- Propenso a sobreajuste em conjuntos de dados ruidosos. - Mais complexo e computacionalmente custoso do que bagging.

Fonte: Do autor (2025).

Os métodos apresentados são adequados para classificar inadimplência em risco de crédito devido à sua combinação de precisão e interpretabilidade. A regressão logística estima probabilidades de inadimplência, enquanto as árvores de decisão identificam fatores-chave de forma clara. Métodos como *bagging* e florestas aleatórias oferecem robustez e reduzem sobreajuste, e o *boosting*, como o XGBoost, entrega alta performance em grandes volumes de dados. Essas características os tornam ideais para prever o risco de crédito com confiabilidade.

Diversos estudos têm empregado modelos de previsão no contexto bancário, especialmente voltados para análise de risco de crédito. Lessmann *et al.* (2015) realizaram um *benchmarking* abrangente de algoritmos de classificação aplicados a crédito, destacando o desempe-

nho superior de modelos baseados em *Random Forest* e *Gradient Boosting*. Malekipirbazari & Aksakalli (2015) utilizaram Random Forest para avaliação de risco em plataformas de empréstimo *peer-to-peer*, obtendo resultados satisfatórios na identificação de inadimplência. No mesmo sentido, Moro, Cortez & Rita (2014) aplicaram técnicas de *data mining* para prever o sucesso de campanhas bancárias, evidenciando a importância da escolha criteriosa dos atributos preditores. Esses trabalhos reforçam a relevância de técnicas de aprendizado de máquina no cenário bancário, sendo base para a proposta deste estudo.

2.10 Trabalhos Relacionados

A regressão logística é amplamente utilizada na modelagem de risco de crédito por sua capacidade de estimar probabilidades de inadimplência com base em variáveis explicativas. Segundo Gonzalez (2018), essa técnica é eficaz para problemas de classificação binária e oferece boa interpretabilidade, sendo especialmente útil em contextos regulatórios.

Ferreira (2024) aplicou esse modelo para avaliar o risco de crédito de microempreendedores, demonstrando que, mesmo com bases de dados limitadas, a abordagem gera resultados consistentes. O estudo reforça a importância da seleção de variáveis relevantes para melhorar a performance preditiva.

Diversos trabalhos empíricos reforçam a aplicabilidade da regressão logística no contexto de crédito. Beserra *et al.* (2022) utilizaram dados do *UCI Machine Learning Repository* para modelar concessão de crédito, obtendo AUC de 0,847 e sensibilidade de 87%. Já Gonçalves, Gouvêa & Mantovani (2013) trabalharam com dados reais de uma instituição financeira, utilizando amostras balanceadas e validação cruzada, o que evidenciou a eficácia do modelo na tomada de decisão bancária. Complementarmente, Vieira (2023) exploraram sua aplicação em conjunto com a Lei do Cadastro Positivo, destacando sua utilidade na mensuração de score de crédito.

Estudos comparativos entre modelos estatísticos e algoritmos de aprendizado de máquina, como *Random Forest* e *XGBoost*, indicam que, embora os métodos de ML frequentemente apresentem maior acurácia, a regressão logística continua sendo preferida em ambientes regulatórios devido à sua transparência e facilidade de interpretação.

Para lidar com bases de alta dimensionalidade e evitar sobreajuste, a penalização via LASSO (*Least Absolute Shrinkage and Selection Operator*) tem sido incorporada à regressão

logística. Sarker (2021) destaca que o LASSO realiza simultaneamente regularização e seleção de variáveis, o que melhora a generalização dos modelos em conjuntos de dados complexos.

Outro desafio recorrente na modelagem de risco de crédito é o desbalanceamento entre classes, uma vez que a maioria dos clientes tende a ser adimplente. Chawla *et al.* (2002) propuseram o método SMOTE (*Synthetic Minority Over-sampling Technique*), que gera exemplos sintéticos da classe minoritária para equilibrar a distribuição e aumentar a capacidade preditiva dos modelos.

Nesse contexto, a escolha das métricas de avaliação torna-se crítica. He & Garcia (2009) recomendam o uso de indicadores como AUC, F1 Score e acurácia balanceada para avaliar modelos em cenários desbalanceados. Essas métricas, derivadas da matriz de confusão, oferecem uma visão mais realista do desempenho do modelo do que a acurácia simples, especialmente quando o objetivo é minimizar falsos negativos.

Esses estudos fundamentam a escolha da regressão logística penalizada com LASSO neste trabalho, bem como a atenção especial ao desbalanceamento das classes e à calibragem do *cut-off*. A presente pesquisa busca aplicar essas técnicas em uma base real de empréstimos e contribuir para a construção de modelos estatísticos mais robustos e interpretáveis.

3 METODOLOGIA

Os dados analisados no artigo foram obtidos do *Kaggle*[1], uma plataforma que oferece uma ampla variedade de conjuntos de dados abrangendo diversos temas.

3.1 Tratamento dos dados

A base de dados utilizada neste estudo contém um total de 32.581 observações, sendo 26.836 clientes adimplentes (82,3%) e 5.745 clientes inadimplentes (17,7%). Ao analisar a distribuição das classes, observa-se um desbalanceamento considerável, com a classe de adimplentes representando uma grande maioria em relação aos inadimplentes. Esse desequilíbrio pode impactar negativamente a performance de modelos de aprendizado de máquina, pois o modelo tende a favorecer a classe majoritária e negligenciar a classe minoritária.

O conjunto de dados, além de apresentar 32.581 observações, contempla variáveis que descrevem tanto o perfil socioeconômico dos clientes quanto as características dos empréstimos contratados. Entre elas, destaca-se a idade (`person_age`), que permite avaliar a faixa etária dos solicitantes, e a renda anual (`person_income`), indicador direto da capacidade financeira. O tempo de emprego (`person_emp_length`) fornece uma medida de estabilidade laboral, enquanto a condição habitacional (`person_home_ownership`) diferencia clientes que possuem imóvel próprio daqueles que vivem em regime de aluguel ou outra situação. No âmbito do crédito, a base inclui o valor solicitado (`loan_amnt`), a taxa de juros aplicada (`loan_int_rate`) e o grau de classificação do empréstimo (`loan_grade`), além da finalidade declarada (`loan_intent`). Complementarmente, constam variáveis relacionadas ao histórico financeiro, como a existência de registros anteriores de inadimplência (`cb_person_default_on_file`) e o tempo de relacionamento com o sistema de crédito (`cb_person_cred_hist_length`). Essa caracterização inicial permite compreender melhor os fatores que podem influenciar o risco de inadimplência e orienta as etapas seguintes de pré-processamento e modelagem.

Após essa caracterização inicial, procedeu-se à etapa de limpeza e pré-processamento dos dados utilizando a linguagem R versão 4.2.1. O Quadro 3.1 apresenta os principais pacotes utilizados, suas respectivas referências e a função desempenhada por cada um no desenvolvimento do trabalho.

[1] https://www.kaggle.com/datasets/laotse/credit-risk-dataset

Quadro 3.1 – Pacotes R utilizados, referências e funções no trabalho

Pacote	**Referência**	**Função no trabalho**
`tidyverse`	Wickham, H. et al. (2019). *Welcome to the tidyverse*. Journal of Open Source Software, 4(43), 1686.	Manipulação e visualização de dados
`dplyr`	Wickham, H., François, R., Henry, L., & Müller, K. (2023). *dplyr: A Grammar of Data Manipulation*. R package version 1.1.3.	Filtragem, agrupamento e transformação de dados
`e1071`	Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2023). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group*. R package version 1.7-13.	Cálculo de estatísticas descritivas e funções auxiliares
`caTools`	Trapletti, A., & Hornik, K. (2023). *caTools: Tools: Moving Window Statistics, GIF, Base64, ROC Curve*. R package version 1.18.2.	Separação dos dados em treino e teste
`caret`	Kuhn, M. (2008). *Building predictive models in R using the caret package*. Journal of Statistical Software, 28(5), 1–26.	Geração da matriz de confusão e avaliação de modelos
`pROC`	Robin, X. et al. (2011). *pROC: an open-source package for R and S+ to analyze and compare ROC curves*. BMC Bioinformatics, 12(1), 77.	Cálculo da curva ROC e AUC
`lmtest`	Zeileis, A., & Hothorn, T. (2002). *Diagnostic Checking in Regression Relationships*. R News, 2(3), 7–10.	Teste qui-quadrado para modelos
`plotly`	Sievert, C. (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC.	Visualização interativa de gráficos
`MLmetrics`	Yan, Y. (2016). *MLmetrics: Machine Learning Evaluation Metrics*. R package version 1.1.1. Disponível em: https://CRAN.R-project.org/package=MLmetrics	Cálculo do F1 Score
`ROCR`	Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). *ROCR: visualizing classifier performance in R*. Bioinformatics, 21(20), 3940–3941.	Avaliação de desempenho de classificadores
`yardstick`	Kuhn, M., Vaughan, D., & Irizarry, R. (2023). *yardstick: Tools for Evaluating Models*. R package version 1.2.0.	Métricas de avaliação de modelos
`ipred`	Peters, A., & Hothorn, T. (2023). *ipred: Improved Predictors*. R package version 0.9-14.	Funções auxiliares para validação cruzada
`grid`	R Core Team. (2023). *grid: The Grid Graphics Package*. R package incluído no base R.	Suporte à construção de gráficos

O tratamento adequado dos dados é essencial para garantir a integridade da análise e evitar viés nos modelos preditivos. A imputação de valores ausentes com médias preserva a estrutura geral da base, enquanto a padronização e renomeação das variáveis facilita a interpretação dos resultados. O desbalanceamento da variável resposta será tratado posteriormente na etapa de modelagem, com estratégias específicas para mitigar seus efeitos.

Este trabalho utiliza uma abordagem computacional em linguagem R para avaliar o desempenho da regressão logística penalizada via LASSO na predição de inadimplência. A seguir, são descritas as etapas de preparação dos dados, modelagem e avaliação, com destaque para os pacotes e funções utilizadas.

3.1.1 Tratamento de valores ausentes e inconsistências

Durante a inspeção inicial, foram identificados valores ausentes em variáveis importantes, como `person_emp_length` (tempo de emprego), com 895 valores ausentes, e `loan_int_rate` (taxa de juros), com 3.116 valores ausentes. Além disso, foram detectados valores inconsistentes, como idades inferiores a 18 anos ou superiores a 100 anos, que foram considerados fora da faixa etária plausível.

As etapas de tratamento aplicadas foram:

- Valores de idade (`person_age`) inferiores a 18 ou superiores a 100 foram substituídos por `NA`, e posteriormente imputados com a média das idades válidas.
- Valores de tempo de emprego (`person_emp_length`) superiores a 100 anos foram tratados como `NA` e imputados com a média dos valores válidos.
- Valores ausentes na taxa de juros (`loan_int_rate`) foram substituídos pela média da variável.

3.1.2 Codificação de variáveis

A variável `cb_person_default_on_file`, originalmente categórica com valores `Y` e `N`, foi recodificada para valores binários: 1 para `Y` (cliente já inadimplente) e 0 para `N` (cliente sem histórico de inadimplência).

3.1.3 Padronização e seleção de variáveis

Após o tratamento, as variáveis foram renomeadas para facilitar a leitura e análise, seguindo convenções mais intuitivas. A base final foi composta pelas seguintes variáveis:

- `default`: variável resposta (inadimplência)
- `taxaJuros`, `vlFinanciado`, `idade`, `renda`, `tempoEmprego`, `tempoRelacionamento`, `comprometimentoDeRenda`
- `objetivoEmprestimo`, `condicaoHabitacional`, `rating`, `historicoDefault`

3.1.4 Construção dos cenários de desbalanceamento

A base original tratada é composta por 32.581 observações. Para permitir a avaliação do comportamento dos modelos em diferentes cenários de tamanho amostral e níveis de desbalanceamento, foram geradas bases sintéticas por meio de procedimentos de reamostragem com reposição, utilizando exclusivamente as observações da base original como fonte de dados.

A partir da base original tratada, foram geradas múltiplas bases sintéticas com diferentes combinações de tamanho amostral e proporção da classe positiva (`default = 1`). Essa etapa foi realizada com a função `gera_df()`, que recebe como parâmetros o número de observações (`n`) e a proporção desejada de inadimplentes (`propDefault`).

Para cada combinação, a função realiza:

- Separação das classes (inadimplentes e adimplentes)
- Amostragem com reposição para atingir o número desejado de registros por classe
- Embaralhamento da base final com `sample_frac(1.0)`

As configurações testadas foram:

- Tamanhos amostrais: 50, 500, 1.000, 10.000, 100.000 e 1.000.000
- Proporções da classe positiva: 20%, 15%, 10%, 5%, 3% e 1%

Ao todo, foram geradas 36 bases sintéticas, todas derivadas exclusivamente da base original por meio de reamostragem com reposição, as quais foram utilizadas nas etapas subsequentes de modelagem e avaliação dos modelos.

Tabela 3.1 – Cenários testados: combinações de tamanho amostral e proporção de inadimplentes

Tamanho amostral (n)	**Proporção de inadimplentes (%)**
50	20, 15, 10, 5, 3, 1
500	20, 15, 10, 5, 3, 1
1.000	20, 15, 10, 5, 3, 1
10.000	20, 15, 10, 5, 3, 1
100.000	20, 15, 10, 5, 3, 1
1.000.000	20, 15, 10, 5, 3, 1

3.1.5 Expansão dos preditores

Para enriquecer o espaço de variáveis e permitir que o modelo penalizado capture relações não lineares e interações, foi criada a função `expand_features()`. Essa função aplica:

- Termos quadráticos e cúbicos: gerados com operações como `num_vars^2` e `num_vars^3`
- Interações duas a duas entre os preditores originais: construídas com `combn()` e `map()` do pacote `purrr`

O resultado é um `tibble` com os preditores originais e suas expansões, que é transformado em matriz com `model.matrix(~ . -1)` para ser compatível com o pacote `glmnet`.

3.1.6 Ajuste do modelo

O modelo utilizado foi a regressão logística penalizada via LASSO, implementada com a função `cv.glmnet()` do pacote `glmnet`. O parâmetro de penalização (`lambda`) foi escolhido automaticamente por validação cruzada com 5 folds, utilizando a métrica AUC como critério de seleção.

A penalização LASSO (`alpha = 1`) tem como característica encolher os coeficientes de variáveis irrelevantes para zero, promovendo uma seleção automática de atributos e reduzindo o risco de *overfitting*, especialmente em cenários com alta dimensionalidade.

3.1.7 Avaliação do Modelo

Após o ajuste, foram geradas as probabilidades de inadimplência para cada observação. Em vez de utilizar um ponto de corte fixo, por exemplo 0,5, foi realizada uma varredura de *cut-offs* entre 0,001 e 0,999. Para cada valor, foi calculada a acurácia balanceada, definida como a média entre sensibilidade e especificidade.

O ponto de corte ótimo foi definido como aquele que maximiza a acurácia balanceada, garantindo um melhor equilíbrio entre a detecção de inadimplentes e a minimização de falsos positivos.

3.1.8 Métricas de desempenho

Para cada base de dados, foram calculadas as principais métricas de avaliação utilizadas em modelos de classificação, incluindo área sob a curva ROC (AUC), F1 Score, acurácia geral, sensibilidade, especificidade, acurácia balanceada e o *cut-off* ótimo. Essas métricas permitiram uma análise abrangente da performance dos modelos em diferentes cenários de desbalanceamento. Além disso, os coeficientes estimados pelo modelo foram extraídos e armazenados, possibilitando a identificação dos efeitos mais relevantes entre as variáveis explicativas.

4 RESULTADOS

Neste capítulo são apresentados os resultados obtidos a partir da aplicação da regressão logística penalizada via LASSO sobre as 36 bases sintéticas geradas com diferentes combinações de tamanho amostral (`n`) e proporção de inadimplentes (`p`). Para cada base, foram calculadas as principais métricas de desempenho, incluindo AUC, F1 Score, acurácia geral, sensibilidade, especificidade, acurácia balanceada e o *cut-off*.

4.1 Desempenho geral dos modelos

De modo geral, os modelos apresentaram desempenho satisfatório em termos de AUC, com valores variando entre aproximadamente 0,83 e 0,99. Observa-se que, mesmo em cenários com forte desbalanceamento (por exemplo, com proporção de inadimplentes de 1%), o ajuste do ponto de corte permitiu alcançar sensibilidade elevada. Em particular, no cenário com 1.000 observações e apenas 1% de inadimplentes, foi possível obter sensibilidade igual a 1,00 com *cut-off* de 0,009.

O F1 Score também apresentou valores elevados na maioria das bases, indicando bom equilíbrio entre precisão e recall. A acurácia balanceada, utilizada como critério para escolha do *cut-off* ótimo, mostrou-se eficaz para compensar o desbalanceamento das classes, com valores superiores a 0,75 mesmo em bases com `p = 0,01`.

4.2 Impacto do desbalanceamento

A análise dos resultados revela que o desbalanceamento da variável resposta impacta diretamente o desempenho dos modelos. Em bases com menor proporção de inadimplentes, a acurácia geral tende a ser mais baixa, mas a sensibilidade pode ser elevada quando o *cut-off* é ajustado corretamente.

Por exemplo, no cenário com 1.000 observações e proporção de inadimplentes de 1%, a acurácia foi de apenas 53%, enquanto a sensibilidade atingiu 100%. Esse resultado indica que o modelo foi capaz de identificar todos os inadimplentes, ainda que à custa de uma elevada taxa de falsos positivos. Em contrapartida, em cenários com distribuições mais equilibradas entre as

classes, como aqueles com proporção de inadimplentes de 20%, observa-se maior estabilidade entre as métricas de desempenho.

4.3 Influência do tamanho amostral

O aumento do tamanho amostral também contribuiu para a estabilidade dos resultados. Bases com 100.000 ou 1.000.000 observações apresentaram métricas mais consistentes, com menor variabilidade entre sensibilidade e especificidade. Isso sugere que, além do ajuste do *cut-off*, o volume de dados é um fator importante para a robustez dos modelos.

4.4 Distribuição dos *cut-offs* ótimos

Os valores de *cut-off* ótimo variaram conforme o grau de desbalanceamento. Em bases com `p = 0,01`, os *cut-offs* ficaram próximos de 0,01, enquanto em bases com `p = 0,2`, os *cut-offs* ótimos se aproximaram de 0,22. Essa variação confirma a importância de ajustar dinamicamente o ponto de decisão, em vez de utilizar um valor fixo como 0,5.

4.5 Desempenho por Configuração Amostral

Nesta seção são apresentados os resultados do desempenho do modelo em diferentes tamanhos amostrais (n) e proporções da classe positiva (inadimplentes). As Tabelas a seguir sintetizam as métricas de avaliação (AUC, F1 Score, Sensibilidade, Especificidade, Acurácia Balanceada e *cut-off* Ótimo) para cada configuração.

A Tabela 4.1 apresenta o desempenho do modelo em amostras muito pequenas ($n = 50$), variando a proporção de inadimplentes. Embora a proporção (p) seja definida, o número absoluto de casos positivos é extremamente baixo (em p=0,01, apenas um inadimplente), o que torna o cenário altamente sensível a perturbações e limita a estabilidade estatística.

Tabela 4.1 – Desempenho do modelo para $n = 50$ com diferentes proporções de inadimplentes

Proporção (p)	**AUC**	**F1 Score**	**Sensibilidade**	**Especificidade**	**Acc. Balanceada**	**Cut-off Ótimo**
0,01	0,898	0,946	1,000	0,898	0,949	0,015
0,03	0,979	0,979	1,000	0,958	0,979	0,380
0,05	0,979	0,979	1,000	0,958	0,979	0,380
0,10	1,000	1,000	1,000	1,000	1,000	0,221
0,15	0,994	0,988	1,000	0,976	0,988	0,209
0,20	0,870	0,861	0,900	0,775	0,838	0,202

A base correspondente à configuração $n = 50$, $p = 0{,}01$ foi incluída na análise após ajustes na geração amostral, garantindo a presença de ao menos um caso da classe positiva (inadimplente). Com essa correção, foi possível calcular métricas como sensibilidade e F1 Score, que dependem da existência de observações positivas. O modelo apresentou sensibilidade igual a 1,00 e acurácia balanceada de 0,949, indicando que o único inadimplente foi corretamente identificado.

No entanto, é importante destacar que, em bases extremamente pequenas e desbalanceadas, pequenas variações nos dados podem gerar grandes oscilações nas métricas, o que limita a generalização dos resultados. Ainda assim, essa configuração demonstra a capacidade do modelo de se ajustar mesmo em cenários com baixa representatividade da classe minoritária.

A Tabela 4.2 mostra os resultados para $n = 500$, ainda em cenários desbalanceados, mas já com maior número absoluto de inadimplentes em cada proporção. Isso permite avaliar se o aumento da amostra reduz a variabilidade das métricas e fornece maior robustez ao modelo, mesmo mantendo a mesma proporção relativa da classe positiva.

Tabela 4.2 – Desempenho do modelo para $n = 500$ com diferentes proporções de inadimplentes

Proporção (p)	**AUC**	**F1 Score**	**Sensibilidade**	**Especificidade**	**Acc. Balanceada**	**Cut-off Ótimo**
0,01	0,859	0,912	0,800	0,840	0,820	0,013
0,03	0,892	0,878	0,867	0,786	0,826	0,033
0,05	0,867	0,873	0,800	0,783	0,792	0,053
0,10	0,918	0,906	0,900	0,838	0,869	0,101
0,15	0,885	0,887	0,813	0,824	0,818	0,156
0,20	0,893	0,858	0,870	0,775	0,823	0,166

Com o aumento do tamanho amostral para $n = 500$, observa-se uma leve redução na variabilidade das métricas em relação à configuração anterior. A sensibilidade permanece elevada,

mesmo em proporções baixas de inadimplentes, e a acurácia balanceada se mantém acima de 0,79 em todos os cenários. Isso indica maior estabilidade do modelo, embora ainda haja impacto do desbalanceamento nas métricas de especificidade.

Na Tabela 4.3, são apresentados os resultados para $n = 1000$. Aqui, o número absoluto de casos positivos cresce significativamente, mesmo em proporções baixas, permitindo observar maior consistência entre as métricas e avaliar a capacidade de generalização do modelo em bases mais robustas.

Tabela 4.3 – Desempenho do modelo para $n = 1000$ com diferentes proporções de inadimplentes

Proporção (p)	**AUC**	**F1 Score**	**Sensibilidade**	**Especificidade**	**Acc. Balanceada**	**Cut-off Ótimo**
0,01	0,836	0,689	1,000	0,525	0,763	0,009
0,03	0,929	0,911	0,900	0,839	0,870	0,035
0,05	0,901	0,892	0,880	0,809	0,845	0,052
0,10	0,901	0,892	0,810	0,822	0,816	0,113
0,15	0,905	0,909	0,807	0,861	0,834	0,207
0,20	0,892	0,885	0,835	0,826	0,831	0,210

Na configuração com $n = 1000$, os resultados mostram maior consistência entre as métricas. A sensibilidade continua alta, especialmente para proporções acima de 3%, e a acurácia balanceada supera 0,83 em quase todos os casos. O modelo demonstra boa capacidade de generalização, mesmo com proporções reduzidas da classe positiva.

A Tabela 4.4 sintetiza o desempenho do modelo em uma base de grande porte ($n = 10000$). Apesar de proporções pequenas de inadimplentes, o número absoluto de casos positivos já é suficiente para reduzir a instabilidade observada em amostras menores, permitindo uma análise mais confiável da calibragem do *cut-off*.

Tabela 4.4 – Desempenho do modelo para $n = 10000$ com diferentes proporções de inadimplentes

Proporção (p)	**AUC**	**F1 Score**	**Sensibilidade**	**Especificidade**	**Acc. Balanceada**	**Cut-off Ótimo**
0,01	0,887	0,908	0,780	0,834	0,807	0,013
0,03	0,862	0,880	0,797	0,790	0,793	0,031
0,05	0,848	0,909	0,706	0,846	0,776	0,070
0,10	0,857	0,885	0,753	0,816	0,784	0,114
0,15	0,858	0,876	0,760	0,812	0,786	0,165
0,20	0,858	0,865	0,760	0,808	0,784	0,216

Com $n = 10000$, o modelo apresenta desempenho robusto e estável. A sensibilidade e a especificidade se equilibram melhor, e o F1 Score permanece elevado. A variação do *cut-off* ótimo acompanha a proporção da classe positiva, refletindo o ajuste dinâmico do ponto de decisão.

A Tabela 4.5 apresenta os resultados para $n = 100000$, destacando como o aumento expressivo da amostra estabiliza as métricas mesmo em cenários de forte desbalanceamento. O maior número absoluto de inadimplentes garante estimativas mais robustas e menos sensíveis a variações pontuais.

Tabela 4.5 – Desempenho do modelo para $n = 100.000$ com diferentes proporções de inadimplentes

Proporção (p)	**AUC**	**F1 Score**	**Sensibilidade**	**Especificidade**	**Acc. Balanceada**	**Cut-off Ótimo**
0,01	0,858	0,877	0,785	0,783	0,784	0,010
0,03	0,860	0,877	0,786	0,787	0,787	0,030
0,05	0,858	0,868	0,787	0,775	0,781	0,048
0,10	0,856	0,891	0,734	0,827	0,780	0,122
0,15	0,857	0,883	0,738	0,827	0,783	0,180
0,20	0,858	0,870	0,749	0,817	0,783	0,227

Na configuração com $n = 100.000$, os resultados mantêm o padrão observado anteriormente. As métricas se estabilizam, e o modelo continua a apresentar bom desempenho mesmo em cenários com apenas 1% de inadimplentes. A acurácia balanceada se mantém próxima de 0,78, evidenciando a eficácia do ajuste do *cut-off*.

Por fim, a Tabela 4.6 mostra o desempenho do modelo em uma base massiva ($n = 1.000.000$). Nesse cenário, mesmo proporções extremamente baixas de inadimplentes resultam em milhares de casos positivos, assegurando máxima estabilidade das métricas e permitindo avaliar a robustez da abordagem em condições de produção.

Tabela 4.6 – Desempenho do modelo para $n = 1.000.000$ com diferentes proporções de inadimplentes

Proporção (p)	**AUC**	**F1 Score**	**Sensibilidade**	**Especificidade**	**Acc. Balanceada**	**Cut-off Ótimo**
0,01	0,853	0,886	0,757	0,797	0,777	0,011
0,03	0,857	0,870	0,785	0,775	0,780	0,029
0,05	0,857	0,898	0,738	0,825	0,782	0,061
0,10	0,858	0,886	0,747	0,817	0,782	0,115
0,15	0,858	0,879	0,748	0,818	0,783	0,171
0,20	0,859	0,870	0,750	0,818	0,784	0,225

Com $n = 1.000.000$ o modelo atinge seu maior nível de estabilidade. As métricas variam pouco entre as diferentes proporções, e o desempenho geral se mantém elevado. A capacidade de identificar inadimplentes com precisão é preservada, mesmo em cenários extremamente desbalanceados, reforçando a robustez da abordagem adotada.

4.5.1 Análise Comparativa dos Resultados

As Tabelas 4.1 a 4.6 apresentam o desempenho do modelo para diferentes tamanhos amostrais (n), variando a proporção de inadimplentes (p). Essa estrutura permite observar com clareza os efeitos do desbalanceamento sobre as métricas de avaliação, especialmente sensibilidade, acurácia balanceada e *Cut-off* ótimo.

Como já discutido nas seções anteriores, o desbalanceamento afeta diretamente a sensibilidade, tornando essencial o ajuste do ponto de corte para garantir a detecção da classe minoritária. A seguir, são destacados os principais padrões observados nas diferentes configurações.

De modo geral, os valores de AUC permaneceram elevados em todas as configurações, variando entre aproximadamente 0,84 e 0,93, o que indica boa capacidade discriminativa do modelo, mesmo em cenários com baixa prevalência da classe positiva.

O F1 Score também apresentou desempenho robusto, com valores superiores a 0,85 na maioria das bases. Isso demonstra equilíbrio entre precisão e recall, especialmente em proporções intermediárias de inadimplentes (entre 5% e 15%), onde o modelo consegue identificar inadimplentes sem penalizar excessivamente os adimplentes.

A sensibilidade mostrou-se altamente influenciada pelo grau de desbalanceamento, especialmente em bases com baixa proporção de inadimplentes. Nas configurações de $n = 50$, o modelo alcançou sensibilidade igual a 1,00 em todas as proporções até $p = 0{,}15$, com leve queda para 0,90 em $p = 0{,}20$. Isso indica que, mesmo em amostras pequenas, o modelo conseguiu identificar todos os inadimplentes em cenários extremamente desbalanceados, desde que o ajuste do *cut-off* fosse adequado. No entanto, essa alta sensibilidade veio acompanhada de variações na especificidade, como no caso de $p = 0{,}01$, em que a especificidade foi de 0,898, evidenciando o *trade-off* entre detecção da classe minoritária e controle de falsos positivos.

Já em $n = 1000$, a configuração com $p = 0{,}01$ também apresentou sensibilidade igual a 1,00, mas com queda acentuada na especificidade (0,525), reforçando esse mesmo padrão. Nas demais bases com $p = 0{,}01$, a sensibilidade variou entre 0,75 e 0,78, indicando que o ajuste do *cut-off* contribui para melhorar a detecção da classe minoritária, mas não garante sensibilidade máxima em todos os cenários. À medida que p aumenta, a sensibilidade tende a se estabilizar em torno de 0,75 a 0,87, com ganhos mais consistentes na acurácia balanceada.

A acurácia balanceada se manteve acima de 0,76 em praticamente todas as configurações, mesmo nas bases mais desbalanceadas. Isso reforça a adequação dessa métrica para avaliar o desempenho em cenários assimétricos, onde a acurácia tradicional pode ser enganosa.

O *cut-off* ótimo variou de forma proporcional ao valor de p. Em bases com $p = 0{,}01$, os valores ficaram próximos de 0,01, enquanto em bases com $p = 0{,}20$, o *cut-off* ideal se aproximou de 0,22. Esse padrão confirma a importância de calibrar dinamicamente o ponto de decisão, evitando o uso de valores fixos como 0,5, que podem comprometer a sensibilidade do modelo em bases desbalanceadas.

Por fim, observa-se que o aumento do tamanho amostral (n) contribuiu para maior estabilidade das métricas. Bases com 100.000 ou 1.000.000 observações apresentaram menor variabilidade entre sensibilidade e especificidade, além de *cut-offs* mais consistentes. Isso sugere que, além do ajuste do *cut-off*, o volume de dados é um fator importante para a robustez da modelagem.

Em síntese, os resultados não apenas confirmam a boa capacidade discriminativa do modelo, mas também evidenciam que a utilidade prática depende da estabilidade das métricas e da calibragem adequada do *cut-off*. Em bases pequenas, a alta sensibilidade é ilusória, pois decorre de poucos casos positivos e não garante generalização. Já em bases maiores, a robustez estatística permite decisões mais confiáveis. Para políticas de crédito, isso significa

que a definição do ponto de corte deve considerar não apenas métricas como AUC ou F1, mas também os *trade-offs* regulatórios, operacionais e econômicos: *cut-offs* conservadores reduzem inadimplência mas limitam inclusão; *cut-offs* permissivos ampliam acesso, mas elevam risco e custos de provisão. Assim, a discussão dos resultados transcende a narrativa dos números e aponta para implicações concretas na gestão de risco e na formulação de políticas de concessão de crédito.

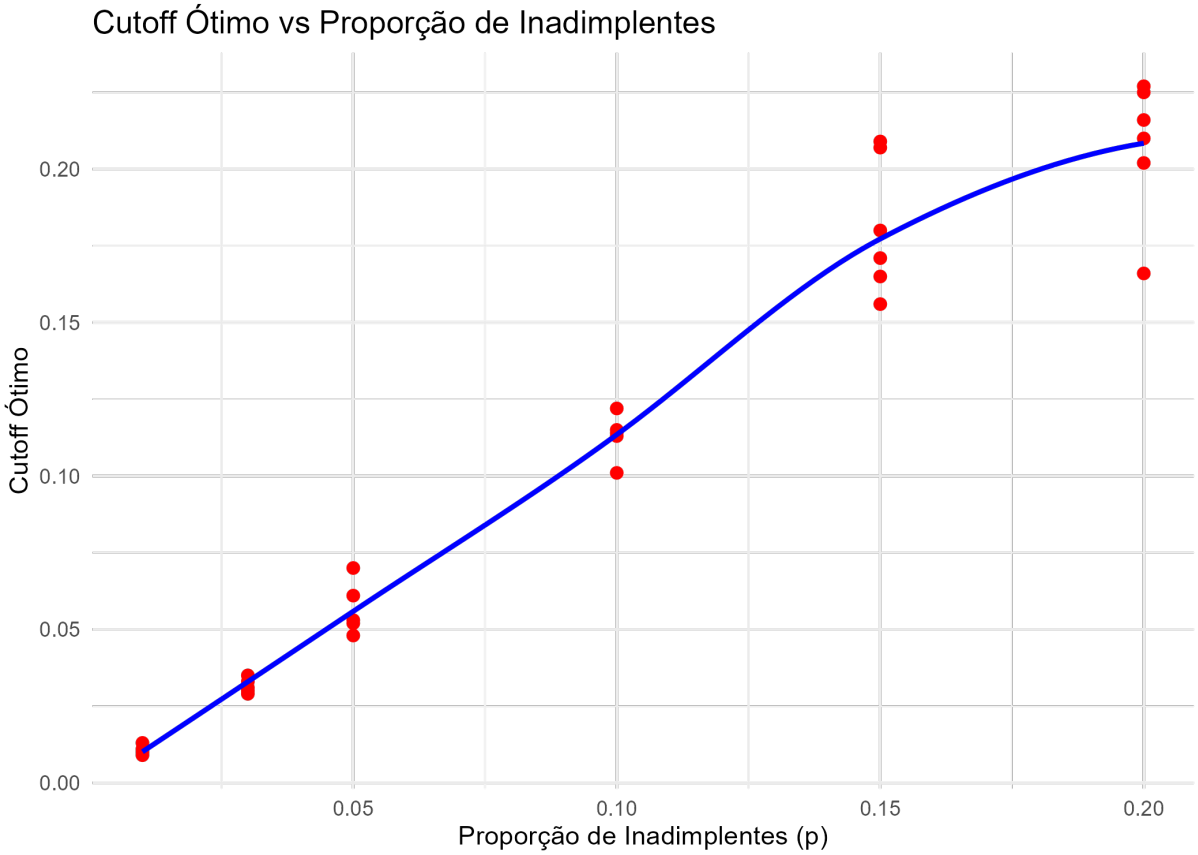

Figura 4.1 – Variação do cut-off ótimo em função da proporção de inadimplentes

A Figura 4.1 apresenta a relação entre a proporção de inadimplentes (`p`) e o valor do *cut-off* ótimo utilizado para maximizar a acurácia balanceada. Observa-se que, conforme o desbalanceamento aumenta (menores valores de `p`), o *cut-off* ideal tende a se aproximar de zero. Isso reforça a importância de calibrar o ponto de decisão conforme o contexto da base, especialmente em cenários com forte assimetria entre as classes.

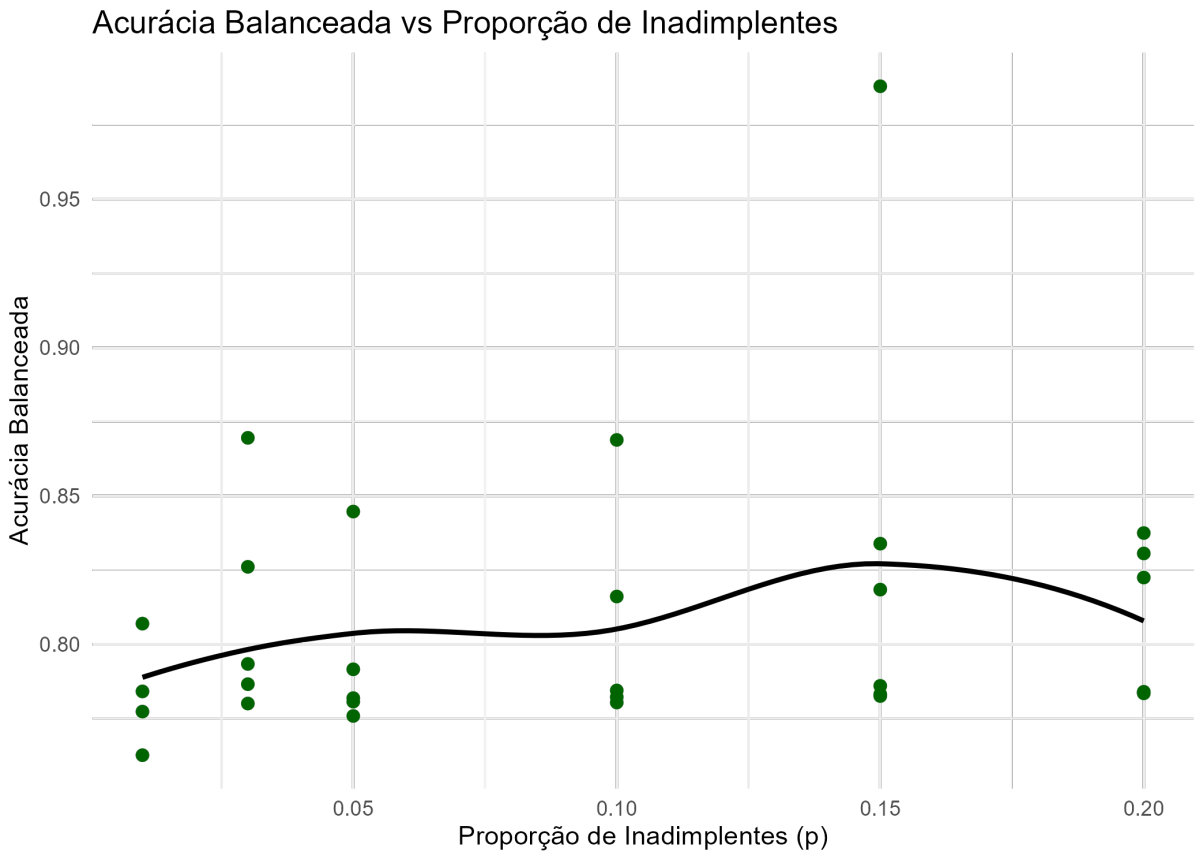

Figura 4.2 – Acurácia balanceada em função da proporção de inadimplentes

A Figura 4.2 mostra como a acurácia balanceada varia conforme a proporção de inadimplentes. Mesmo em bases com `p = 0,01`, o modelo conseguiu manter valores elevados de acurácia balanceada, evidenciando a eficácia do ajuste do *cut-off* e da penalização LASSO na detecção da classe minoritária.

Tabela 4.7 – Principais coeficientes estimados pelo modelo logístico penalizado (LASSO) — cenário com 100.000 observações e 1% de inadimplentes

Variável	**Coeficiente (β)**	**Interpretação**
(Intercept)	-6,197	Probabilidade base de inadimplência é baixa
vlFinanciado	0,00096	Financiamentos maiores aumentam o risco
historicoDefault	0,158	Histórico de inadimplência eleva o risco
comprometimentoDeRenda	9,725	Alto comprometimento de renda aumenta fortemente o risco
taxaJuros_2	0,016	Juros elevados contribuem para inadimplência
tempoEmprego_2	0,0088	Tempo de emprego tem leve efeito positivo
comprometimentoDeRenda_2	52,445	Efeito quadrático acentuado no risco
comprometimentoDeRenda_3	-61,933	Efeito cúbico sugere reversão em altos níveis
taxaJuros_x_comprometimentoDeRenda	-0,363	Juros altos com renda comprometida reduzem risco marginal
vlFinanciado_x_comprometimentoDeRenda	-0,00022	Financiamento alto com renda comprometida reduz risco marginal
historicoDefault_x_comprometimentoDeRenda	-1,523	Interação negativa: histórico ruim com renda comprometida reduz risco marginal

A Tabela 4.7 apresenta os principais coeficientes estimados pelo modelo logístico penalizado via LASSO. Observa-se que variáveis como `comprometimentoDeRenda`, `vlFinanciado` e `historicoDefault` possuem impacto significativo na probabilidade de inadimplência. A presença de termos quadráticos e cúbicos indica efeitos não lineares, enquanto as interações revelam como combinações específicas de variáveis influenciam o risco. Por se tratar de um modelo penalizado, os coeficientes foram selecionados automaticamente com base em sua relevância preditiva, não sendo fornecidos valores-p tradicionais.

Tabela 4.8 – Coeficientes estimados pelo modelo logístico penalizado (LASSO) — cenário com 10.000 observações e proporção de inadimplentes de 5%

Variável	**Coeficiente (β)**	**Interpretação**
(Intercept)	-4,964	Probabilidade base de inadimplência é baixa.
vlFinanciado	0,00055	Financiamentos maiores aumentam o risco.
tempoRelacionamento	-0,033	Relações mais longas reduzem o risco.
historicoDefault	0,140	Histórico de inadimplência eleva o risco.
comprometimentoDeRenda	11,544	Alto comprometimento de renda aumenta fortemente o risco.
taxaJuros_2	0,014	Juros elevados contribuem para inadimplência.
comprometimentoDeRenda_2	29,533	Efeito quadrático acentuado no risco.
comprometimentoDeRenda_3	-33,710	Efeito cúbico sugere reversão em níveis muito altos.
taxaJuros_x_comprometimentoDeRenda	-0,267	Juros altos com renda comprometida reduzem o risco marginal.
vlFinanciado_x_tempoRelacionamento	0,0000053	Relações mais longas com financiamento elevado aumentam levemente o risco.
idade_x_comprometimentoDeRenda	-0,0586	Idade mais alta com renda comprometida reduz o risco marginal.
historicoDefault_x_comprometimentoDeRenda	-2,192	Interação negativa: histórico ruim com renda comprometida reduz o risco marginal.

A Tabela 4.8 apresenta os coeficientes estimados pelo modelo logístico penalizado via LASSO para a base com 10.000 observações e 5% de inadimplentes. Observa-se que variáveis como `vlFinanciado`, `comprometimentoDeRenda` e `historicoDefault` foram selecionadas como preditores relevantes, com sinais coerentes com a literatura: financiamentos mais altos e histórico de inadimplência aumentam o risco, enquanto maior tempo de relacionamento com a instituição financeira tende a reduzi-lo.

Além dos efeitos principais, o modelo identificou termos quadráticos e cúbicos, como `comprometimentoDeRenda_2` e `comprometimentoDeRenda_3`, indicando que o impacto da renda comprometida sobre o risco não é linear — há intensificação do risco em níveis moderados e possível reversão em níveis extremos. Isso reforça a importância da expansão de variáveis para capturar padrões complexos.

As interações também revelam nuances importantes. Por exemplo, a combinação de `taxaJuros` com `comprometimentoDeRenda` apresentou efeito negativo, sugerindo que, em certos contextos, juros elevados podem atenuar o risco marginal quando a renda já está comprometida. A interação entre `historicoDefault` e `comprometimentoDeRenda` também mostrou efeito negativo, o que pode indicar que clientes com histórico negativo e alta renda comprome-

tida já estão em um perfil de risco extremo, onde o modelo ajusta a probabilidade com mais cautela.

Esses resultados demonstram que, mesmo em bases com desbalanceamento, a regressão logística penalizada é capaz de selecionar variáveis relevantes e modelar relações não lineares e interativas, oferecendo uma estrutura interpretável e eficaz para a análise de risco de crédito.

5 DISCUSSÃO DOS RESULTADOS

A análise dos resultados obtidos com a regressão logística penalizada via LASSO permite discutir não apenas o desempenho preditivo dos modelos, mas também suas implicações metodológicas e práticas no contexto da modelagem de risco de crédito em bases desbalanceadas. Diferentemente de uma análise puramente técnica, os achados deste estudo dialogam diretamente com limitações já apontadas na literatura e oferecem evidências empíricas que reforçam, e em alguns casos refinam, resultados previamente estabelecidos.

Um dos principais pontos observados diz respeito à forte sensibilidade dos modelos ao desbalanceamento da variável resposta. Conforme amplamente discutido por He & Garcia (2009) e Chawla *et al.* (2002), classificadores treinados em bases altamente assimétricas tendem a apresentar desempenho ilusório quando avaliados por métricas tradicionais, como a acurácia global. Os resultados deste trabalho confirmam esse comportamento: em cenários com baixa proporção de inadimplentes, modelos avaliados com *cut-off* fixo apresentaram elevada acurácia aparente, mas baixa capacidade de identificação da classe minoritária.

Entretanto, um avanço importante observado neste estudo está na demonstração empírica de que o ajuste dinâmico do *cut-off*, aliado à utilização da acurácia balanceada como critério de otimização, permite recuperar sensibilidade elevada mesmo em cenários extremos de desbalanceamento. Esse resultado vai além de simplesmente confirmar a literatura: ele evidencia que, mesmo sem recorrer explicitamente a técnicas de reamostragem voltadas à geração de novas observações sintéticas, é possível obter modelos eficazes desde que o processo decisório seja devidamente calibrado. Tal achado dialoga com abordagens como o SMOTE (Chawla *et al.*, 2002), apresentando-se como uma alternativa metodológica mais simples e interpretável.

Embora os experimentos tenham sido conduzidos em bases sintéticas obtidas por reamostragem com reposição, as implicações desses resultados são particularmente relevantes para aplicações em bases reais. Em ambientes regulados, como o risco de crédito, o uso de dados sintéticos pode ser questionado por auditores e órgãos reguladores, que frequentemente demandam decisões baseadas em dados observados. Nesse contexto, a calibragem do *cut-off* emerge como uma estratégia atrativa, por atuar no processo decisório sem alterar a estrutura original dos dados.

Outro aspecto relevante diz respeito ao impacto do tamanho amostral na estabilidade dos modelos. Trabalhos anteriores apontam que métodos penalizados tendem a apresentar maior ro-

bustez à medida que o volume de dados aumenta (Tibshirani, 1996). Os resultados aqui obtidos corroboram essa afirmação: bases com 100.000 ou mais observações apresentaram menor variabilidade entre sensibilidade e especificidade, além de *cut-offs* mais consistentes. Isso sugere que, em contextos operacionais reais, onde grandes volumes de dados estão disponíveis, a regressão logística penalizada via LASSO constitui uma ferramenta particularmente adequada.

Os resultados deste estudo estão alinhados com a literatura que defende o uso de métodos penalizados em contextos de modelagem de risco de crédito, especialmente diante de alta dimensionalidade e desbalanceamento de classes. A regressão logística penalizada via LASSO, proposta por Tibshirani (1996), mostrou-se eficaz ao realizar simultaneamente seleção de variáveis e regularização dos coeficientes, permitindo a inclusão de termos não lineares e interações sem comprometer a estabilidade do modelo. Esse comportamento reforça evidências apresentadas por Hastie, Tibshirani & Friedman (2009b), segundo as quais métodos penalizados tendem a ser mais robustos do que procedimentos tradicionais de seleção, como o *stepwise*, que são conhecidos por sua instabilidade e sensibilidade a variações amostrais.

A análise dos coeficientes selecionados pelo LASSO também revelou padrões consistentes com a literatura de risco de crédito. Variáveis como comprometimento de renda, valor financiado e histórico de inadimplência figuram entre os principais determinantes do risco, em consonância com estudos clássicos da área. No entanto, um diferencial deste trabalho está na identificação sistemática de efeitos não lineares e interações relevantes, como a reversão do efeito do comprometimento de renda em níveis extremos e a modulação do risco pela combinação entre taxa de juros e perfil financeiro do cliente. Esses resultados reforçam achados mais recentes que defendem a incorporação de relações não lineares na modelagem de crédito, mesmo em estruturas estatísticas relativamente simples.

Do ponto de vista metodológico, a escolha do LASSO mostrou-se particularmente vantajosa. Em comparação com procedimentos tradicionais, como o *stepwise*, o LASSO oferece maior estabilidade na seleção de variáveis, menor risco de sobreajuste e melhor desempenho preditivo em cenários de alta dimensionalidade e multicolinearidade (Tibshirani, 1996). Além disso, diferentemente de métodos puramente preditivos de caixa-preta, a regressão logística penalizada preserva a interpretabilidade dos coeficientes, aspecto fundamental em aplicações financeiras sujeitas a auditoria e regulamentação. Embora métodos alternativos pudessem ser empregados, os resultados indicam que o LASSO representa um compromisso eficaz entre desempenho, robustez e transparência.

As implicações práticas desses achados são diretas. O estudo demonstra que decisões críticas em risco de crédito não devem se apoiar em *cut-offs* fixos nem em métricas inadequadas ao contexto dos dados. A calibragem do ponto de decisão emerge como uma ferramenta estratégica, capaz de alinhar objetivos estatísticos e operacionais. Além disso, a evidência de estabilidade do modelo em grandes amostras reforça sua aplicabilidade em sistemas reais de concessão e monitoramento de crédito.

Em síntese, este trabalho contribui para a literatura ao fornecer evidências empíricas de que a combinação entre regressão logística penalizada, engenharia de atributos e ajuste dinâmico do *cut-off* constitui uma abordagem robusta, interpretável e eficaz para a modelagem de risco de crédito em bases desbalanceadas. Mais do que confirmar resultados existentes, os achados ampliam a compreensão sobre como decisões aparentemente técnicas, como a escolha do *cut-off*, podem ter impacto substantivo na performance e na utilidade prática dos modelos.

6 CONCLUSÃO

Este trabalho teve como objetivo avaliar o desempenho da regressão logística penalizada via LASSO na predição de inadimplência, considerando diferentes cenários de desbalanceamento e tamanhos amostrais. Foram geradas 36 bases sintéticas com proporções variadas de inadimplentes, e aplicadas técnicas de expansão de preditores, validação cruzada e ajuste dinâmico do ponto de corte.

Os resultados demonstraram que o ajuste do *cutoff* é essencial para lidar com o desbalanceamento da variável resposta. A acurácia balanceada mostrou-se uma métrica eficaz para avaliar o desempenho dos modelos em bases assimétricas, permitindo identificar inadimplentes mesmo em cenários com baixa prevalência da classe positiva.

Além disso, observou-se que o aumento do tamanho amostral contribui para maior estabilidade das métricas, reforçando a importância de bases robustas em aplicações reais. A penalização LASSO foi eficaz na seleção automática de variáveis, reduzindo a complexidade do modelo e mantendo os efeitos mais relevantes.

Este estudo destacou o potencial da regressão logística penalizada, frequentemente considerada um modelo simples, ao mostrar que, com ajustes criteriosos de *cutoff* e o uso de métricas adequadas, como Acurácia Balanceada, F1 Score e AUC, é possível alcançar resultados robustos e interpretáveis na predição de risco de crédito. Assim, mesmo sem recorrer a modelos mais complexos, obteve-se desempenho consistente, aliado à vantagem da interpretabilidade e da conformidade regulatória.

Como contribuição prática, o trabalho oferece uma abordagem sistemática para calibrar modelos preditivos em contextos de crédito, com ênfase na escolha do *cutoff*, na engenharia de atributos e na seleção de métricas apropriadas. Em aplicações reais, por exemplo, um *cutoff* ajustado abaixo de 0,5 pode reduzir perdas por rejeição indevida em cenários altamente desbalanceados, enquanto ajustes finos em bases mais equilibradas podem evitar a aprovação de clientes com alto risco, protegendo a instituição contra inadimplência futura.

Esse tipo de calibragem, orientada por métricas robustas, permite que o modelo se adapte ao perfil da carteira analisada, oferecendo suporte mais preciso à tomada de decisão. Em um cenário financeiro cada vez mais orientado por dados, essa flexibilidade representa um diferencial competitivo e operacional.

Dessa forma, os resultados obtidos permitem responder aos objetivos propostos neste trabalho. A análise exploratória e o pré-processamento das bases sintéticas foram fundamentais para garantir a consistência dos dados. O desempenho dos modelos foi avaliado por meio de métricas adequadas ao contexto de desbalanceamento, evidenciando a capacidade da regressão logística penalizada em lidar com esse tipo de desafio. Além disso, a investigação sobre o impacto do desbalanceamento e o ajuste do *cutoff* mostrou-se essencial para melhorar a sensibilidade e a estabilidade dos modelos, atendendo aos objetivos específicos estabelecidos.

Cabe destacar que a estrutura do trabalho privilegiou a transparência metodológica e a reprodutibilidade dos experimentos, com detalhamento das etapas de simulação, ajuste e avaliação dos modelos. Essa opção, embora possa conferir ao texto um caráter mais didático em alguns trechos, foi adotada de forma deliberada, visando garantir a replicação dos resultados em diferentes contextos e áreas de aplicação, como saúde, segurança pública e finanças. Assim, o foco não esteve apenas nos resultados empíricos, mas na consolidação de uma abordagem sistemática e reutilizável para o tratamento de problemas de classificação com desbalanceamento.

Apesar dos resultados promissores, este estudo apresenta algumas limitações. A análise foi conduzida exclusivamente com bases sintéticas, escolha metodológica que permitiu controlar rigorosamente o grau de desbalanceamento e o tamanho amostral, embora isso possa limitar a generalização direta para dados reais, que apresentam maior variabilidade e ruído. Além disso, foi adotado apenas o modelo de regressão logística penalizada via LASSO, sem comparação com métodos mais sofisticados, como modelos não lineares ou técnicas de *ensemble*. Essa escolha visou preservar a interpretabilidade e o foco metodológico, mas abre espaço para investigações complementares.

Diante desses resultados, abrem-se caminhos promissores para trabalhos futuros, entre os quais destacam-se:

- Aplicar o modelo a bases reais de instituições financeiras, avaliando restrições regulatórias e requisitos de explicabilidade;
- Comparar o desempenho com outras penalizações, como Elastic Net;
- Explorar modelos não lineares, como XGBoost ou redes neurais;
- Investigar estratégias de reamostragem, como SMOTE ou ROSE.

Em síntese, os achados reforçam que, mesmo com modelos estatisticamente simples, é possível obter resultados robustos quando há cuidado na preparação dos dados, na escolha das métricas e na calibração dos parâmetros.

REFERÊNCIAS

ANDRADE, M.; SILVA, R. Modelo de classificação de risco de crédito de empresas. **Revista Contabilidade e Finanças**, 2021. Disponível em: https://www.scielo.br/j/rcf/a/bqnyg6hByqQHY3dyPzm7b5q/?format=html.

Assaf Neto, A. **Mercado financeiro**. 14. ed. São Paulo: Atlas, 2018.

ASSUNÇÃO, G. O.; IZBICKI, R.; PRATES, M. O. Is augmentation effective in improving prediction in imbalanced datasets? **Journal of Data Science**, School of Statistics, Renmin University of China, p. 1–16, 2024. ISSN 1680-743X.

Banco Central do Brasil. **Gestão Integrada de Riscos no Banco Central do Brasil**. 2023. https://www.bcb.gov.br/htms/getriscos/Gestao-Integrada-de-Riscos.pdf. Documento institucional.

BERGSTRA, J.; BENGIO, Y. Random search for hyper-parameter optimization. **Journal of Machine Learning Research**, v. 13, n. Feb, p. 281–305, 2012.

BESERRA, R. S. *et al.* Modelagem com regressão logística para análise de concessão de crédito. **Research, Society and Development**, v. 11, n. 7, 2022. Acesso em: 08 set. 2025. Disponível em: https://rsdjournal.org/index.php/rsd/article/view/29761.

BORDIGNON, A. W. B. **Algoritmo Genético Multiobjetivo: uma aplicação para seleção de portfólios de crédito**. Dissertação (Dissertação de Mestrado) — Escola de Economia de São Paulo, Fundação Getulio Vargas, São Paulo, 2020. Área de Concentração: Engenharia Financeira.

BRITO, G. A. S.; NETO, A. A. Modelo de classificação de risco de crédito de empresas. **Revista Contabilidade amp; Finanças**, v. 19, n. 46, p. 18–29, abr. 2008. Disponível em: https://revistas.usp.br/rcf/article/view/34249.

BRODERSEN, K. H. *et al.* The balanced accuracy and its posterior distribution. In: **2010 20th International Conference on Pattern Recognition**. [*S.l.: s.n.*], 2010. p. 3121–3124.

BUSSMANN, N. *et al.* Explainable machine learning in credit risk management. **Computational Economics**, v. 57, n. 1, p. 203–216, 2021. ISSN 1572-9974. Disponível em: https://doi.org/10.1007/s10614-020-10042-0.

CHAWLA, N. V. *et al.* Smote: Synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**, AI Access Foundation, v. 16, p. 321–357, jun. 2002. ISSN 1076-9757. Disponível em: http://dx.doi.org/10.1613/jair.953.

COELHO, F. F.; AMORIM, D. P. d. L.; CAMARGOS, M. A. d. Analisando métodos de machine learning e avaliação do risco de crédito. **Revista Gestão & amp; Tecnologia**, v. 21, n. 1, p. 89–116, mar. 2021. Disponível em: https://revistagt.fpl.emnuvens.com.br/get/article/view/2089.

DUBEY, R. *et al.* Analysis of sampling techniques for imbalanced data: An n=648 adni study. **NeuroImage**, v. 87, p. 220–241, 2014. ISSN 1053-8119. Disponível em: https://www.sciencedirect.com/science/article/pii/S1053811913010161.

ENCORD. **F1 Score in Machine Learning: What It Is and How to Use It**. 2023. https://encord.com/blog/f1-score-in-machine-learning/. Acesso em: 08 set. 2025.

FARIAS, I. da S. **Ciência de Dados no Mercado de Crédito: Estratégias para Mitigação de Riscos**. 2023. https://ric.cps.sp.gov.br/bitstream/123456789/19765/3/informaticanegocios_2023_2_isaacdasilvafarias_cienciadedadosnomercadodecreditoestrategiaspara.pdf. Centro Paula Souza.

FAYZ, S.; RIZKA, M.; MAGHRABY, F. Cervical cancer diagnosis using random forest classifier with smote and feature reduction techniques. **IEEE Access**, PP, p. 1–1, 10 2018.

FERNáNDEZ, A. *et al.* Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. **J. Artif. Int. Res.**, AI Access Foundation, El Segundo, CA, USA, v. 61, n. 1, p. 863–905, jan. 2018. ISSN 1076-9757.

FERREIRA, G. M. T. **Regressão logística aplicada à gestão de risco de crédito**. 2024. Projeto de Graduação, Universidade Federal do Rio de Janeiro. Disponível em: http://www.repositorio.poli.ufrj.br/monografias/projpoli10043565.pdf.

FUADAH, Y. N.; PRAMUDITO, M. A.; LIM, K. M. An optimal approach for heart sound classification using grid search in hyperparameter optimization of machine learning. **Bioengineering**, v. 10, n. 1, 2023. ISSN 2306-5354. Disponível em: https://www.mdpi.com/2306-5354/10/1/45.

FURRIEL, W. O. **Avaliação do impacto de conjuntos de dados desbalanceados em modelos de classificação para risco de crédito**. 2023. https://acervodigital.ufpr.br/xmlui/bitstream/handle/1884/85723/R%20-%20E%20-%20WESLEY%20OLIVEIRA%20FURRIEL.pdf. Monografia - UFPR.

FÁVERO, L. P. L. *et al.* **Análise de dados: modelagem multivariada para tomada de decisões**. [*S.l.*]: Elsevier, 2009.

GONZALEZ, L. A. **Regressão Logística e suas Aplicações**. 2018. Trabalho de Conclusão de Curso, Universidade Federal do Maranhão. Disponível em: https://monografias.ufma.br/jspui/bitstream/123456789/3572/1/LEANDRO-GONZALEZ.pdf.

GONçALVES, E. B.; GOUVêA, M. A.; MANTOVANI, D. M. N. Análise de risco de crédito com o uso de regressão logística. **Revista Contemporânea de Contabilidade**, v. 10, n. 20, p. 139–160, 2013. Acesso em: 08 set. 2025. Disponível em: https://doi.org/10.5007/2175-8069.2013v10n20p139.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2nd. ed. [*S.l.*]: Springer, 2009.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2. ed. New York: Springer, 2009.

HE, H.; GARCIA, E. A. Learning from imbalanced data. **IEEE Transactions on Knowledge and Data Engineering**, v. 21, n. 9, p. 1263–1284, 2009.

HOSMER, D.; LEMESHOW, S.; STURDIVANT, R. **Applied Logistic Regression**. [*S.l.*]: Wiley, 2013. (Wiley Series in Probability and Statistics). ISBN 9780470582473.

IZBICKI, R.; SANTOS, T. M. dos. **Aprendizado de máquina: uma abordagem estatística**. [*S.l.: s.n.*], 2020. ISBN 978-65-00-02410-4.

JABÔR, R. M. **Administração Ativa de Portfólio de Crédito a Pessoa Jurídica: uma revisão dos principais conceitos para uma implementação efetiva em bancos comerciais**. 84 p. Dissertação (Dissertação de Mestrado) — Escola de Economia de São Paulo, Fundação Getulio Vargas, São Paulo, 2006. Orientador: João Carlos Douat.

JAMES, G. *et al.* **An Introduction to Statistical Learning: with Applications in R**. Springer, 2013. Disponível em: https://faculty.marshall.usc.edu/gareth-james/ISL/.

JAPKOWICZ, N.; STEPHEN, S. The class imbalance problem: A systematic study. **Intell. Data Anal.**, IOS Press, NLD, v. 6, n. 5, p. 429–449, out. 2002. ISSN 1088-467X.

LESSMANN, S. *et al.* Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. **European Journal of Operational Research**, v. 247, n. 1, p. 124–136, 2015. ISSN 0377-2217. Disponível em: https://www.sciencedirect.com/science/article/pii/S0377221715004208.

MALEKIPIRBAZARI, M.; AKSAKALLI, V. Risk assessment in social lending via random forests. **Expert Systems with Applications**, 02 2015.

MILLER, C. *et al.* A review of model evaluation metrics for machine learning in genetics and genomics. **Frontiers in Bioinformatics**, v. 4, 2024. ISSN 2673-7647. Disponível em: https://www.frontiersin.org/journals/bioinformatics/articles/10.3389/fbinf.2024.1457619.

MORO, S.; CORTEZ, P.; RITA, P. A data-driven approach to predict the success of bank telemarketing. **Decision Support Systems**, v. 62, p. 22–31, 2014. ISSN 0167-9236. Disponível em: https://www.sciencedirect.com/science/article/pii/S016792361400061X.

NETO, R. F. M. **Análise de Sensibilidade dos Modelos KMV, de Merton, e CreditRisk+ de Gestão de Portfólio de Crédito**. x, 108 p. Dissertação (Dissertação de Mestrado) — Universidade Presbiteriana Mackenzie / Centro de Ciências Sociais Aplicadas, São Paulo, 2011. Orientador: Herbert Kimura; Programa de Pós-Graduação em Administração.

RAMEZAN, C. A.; WARNER, T. A.; MAXWELL, A. E. Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification. **Remote Sensing**, v. 11, n. 2, 2019. ISSN 2072-4292. Disponível em: https://www.mdpi.com/2072-4292/11/2/185.

RUPAPARA, V. *et al.* Impact of smote on imbalanced text features for toxic comments classification using rvvc model. **IEEE Access**, PP, p. 1–1, 05 2021.

SARKER, I. H. Machine learning: Algorithms, real-world applications and research directions. **SN Comput. Sci.**, Springer-Verlag, Berlin, Heidelberg, v. 2, n. 3, mar. 2021. Disponível em: https://doi.org/10.1007/s42979-021-00592-x.

SAUNDERS, A.; ALLEN, L. **Credit Risk Measurement: New Approaches to Value at Risk and Other Paradigms**. Wiley, 2002. (Wiley Finance). ISBN 9780471274766. Disponível em: https://books.google.com.br/books?id=pGMLAd2WvasC.

SCHUSTER, W. E.; REIS, M. d. Otimizando a alocaÇÃo de capital: AnÁlise do setor bancÁrio por meio do modelo raroc. **Contabilidade Vista amp; Revista**, v. 35, n. 2, p. 104–126, dez. 2024. Disponível em: https://revistas.face.ufmg.br/index.php/contabilidadevistaerevista/article/view/8080.

SECURATO, J. R. **Crédito: Análise e avaliação do risco**. 2. ed. São Paulo: Saint Paul, 2015.

StoneX. **O que é risco de crédito? Significado, papel e exemplos**. 2023. https://www.stonex.com/pt-br/glossario-financeiro/risco-de-credito/. Artigo institucional.

TAN, X. *et al.* Wireless sensor networks intrusion detection based on smote and the random forest algorithm. **Sensors**, v. 19, n. 1, 2019. ISSN 1424-8220. Disponível em: https://www.mdpi.com/1424-8220/19/1/203.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley, v. 58, n. 1, p. 267–288, 1996.

VIEIRA, L. **Risco de crédito e a aplicação da modelagem regressão logística**. Dissertação (Dissertação de Mestrado) — Universidade Estadual Paulista "Júlio de Mesquita Filho" – UNESP, 2023. Acesso em: 08 set. 2025. Disponível em: https://repositorio.unesp.br/bitstreams/92638ccd-bd0d-427a-a651-abfc73b134f3/download.

VUJOVIĆ, Ž. *et al.* Classification model evaluation metrics. **International Journal of Advanced Computer Science and Applications**, v. 12, n. 6, p. 599–606, 2021.

XU, Z. *et al.* Class-weighted classification: Trade-offs and robust approaches. In: III, H. D.; SINGH, A. (Ed.). **Proceedings of the 37th International Conference on Machine Learning**. PMLR, 2020. (Proceedings of Machine Learning Research, v. 119), p. 10544–10554. Disponível em: https://proceedings.mlr.press/v119/xu20b.html.

YIN, J.; HAN, B.; WONG, H. Y. Covid-19 and credit risk: A long memory perspective. **Insurance: Mathematics and Economics**, v. 104, p. 15–34, 2022. ISSN 0167-6687. Disponível em: https://www.sciencedirect.com/science/article/pii/S0167668722000142.