

Ciência de Dados no Mercado de Crédito: Estratégias para Mitigação de Riscos e Otimização de Decisões com modelagem preditiva

Isaac da Silva Farias, Marildo Domingos da Silva
e-mail: isaac.farias@fatec.sp.gov.br; marildo.silva@fatec.sp.gov.br;

Faculdade de Tecnologia de São José do Rio Preto

Resumo: O setor financeiro enfrenta desafios significativos no cenário do mercado de crédito, com foco especial na concessão de crédito para pessoa física. O risco de crédito, notadamente a inadimplência, é uma ameaça central, exigindo estratégias de gerenciamento eficazes. Modelos tradicionais baseados em histórico de crédito, capacidade de pagamento e garantias são cruciais, mas a má gestão financeira é identificada como uma causa substancial de insolvência.

Taxas de juros elevadas, fraude e regulamentação também compõem os desafios enfrentados pelo mercado. A inadimplência não apenas resulta em perdas financeiras diretas para os credores, mas também impacta a rentabilidade e a reputação. A má gestão financeira contribui para custos elevados de recuperação, restrições de recursos para empréstimos futuros e ameaça a reputação do credor.

A ciência de dados e a estatística emergem como soluções poderosas para enfrentar esses desafios. Modelos preditivos, análises avançadas e algoritmos de aprendizado de máquina proporcionam insights valiosos. A análise em tempo real permite decisões de crédito, otimização de recursos e prevenção eficaz de fraudes.

A regulamentação desempenha um papel crítico na garantia de práticas justas e seguras, enquanto a ciência de dados oferece uma abordagem proativa, promovendo educação financeira e práticas de concessão de crédito responsáveis. Em resumo, a aplicação eficaz de modelos preditivos e análises estatísticas no mercado de crédito para pessoa física representa um avanço crucial para mitigar riscos, melhorar eficiência e assegurar conformidade regulatória.

Palavras-chave: Setor financeiro, Risco de crédito, Insolvência, Ciência de dados, Algoritmos de aprendizado de máquina,

Abstract: *The financial sector faces significant challenges in the credit market landscape, with a special focus on granting credit to individuals. Credit risk, notably delinquency, stands as a central threat, necessitating effective management strategies. Traditional models based on credit history, repayment capacity, and collateral are crucial, but poor financial management is identified as a substantial cause of insolvency.*

High-interest rates, fraud, and regulation also constitute challenges encountered by the market. Delinquency not only results in direct financial losses for creditors but also impacts profitability and reputation. Poor financial management contributes to high recovery costs, resource constraints for future loans, and threatens the lender's reputation.

Data science and statistics emerge as powerful solutions to tackle these challenges. Predictive models, advanced analytics, and machine learning algorithms provide valuable insights. Real-time analysis enables credit decisions grounded in statistical evidence, resource optimization, and effective fraud prevention.

Regulation plays a critical role in ensuring fair and secure practices, while data science offers a proactive approach, promoting financial education and responsible credit-granting practices. In summary, the effective application of predictive models and statistical analyses in the credit market for individuals represents a crucial advancement in mitigating risks, improving efficiency, and ensuring regulatory compliance.

Keywords: *Financial sector, Credit risk, Insolvency, Data science, Machine learning algorithms,*

1 INTRODUÇÃO

No dinâmico cenário do setor financeiro, o mercado de crédito para pessoa física enfrenta desafios substanciais, demandando abordagens inovadoras e estratégias eficazes de gerenciamento. Entre os desafios, destaca-se o risco de crédito, com foco especial na inadimplência, cujas implicações impactam não apenas a estabilidade financeira das instituições,

mas também sua reputação no mercado. Este artigo explora os desafios enfrentados pelo mercado de crédito, incluindo taxas de juros elevadas, fraudes e regulamentações rigorosas, enquanto destaca a má gestão financeira como uma causa significativa de insolvência.

Diante desses desafios, surgem oportunidades para a aplicação eficaz de ciência de dados e estatística como ferramentas poderosas na mitigação de riscos, aprimoramento da eficiência operacional e garantia de conformidade regulatória. Modelos preditivos, análises avançadas e algoritmos de aprendizado de máquina emergem como soluções promissoras, proporcionando insights valiosos para informar decisões de crédito, otimizar a alocação de recursos e prevenir eficazmente fraudes.

Além disso, a regulamentação desempenha um papel crítico na promoção de práticas justas e seguras no mercado de crédito, enquanto a ciência de dados oferece uma abordagem proativa ao promover a educação financeira e práticas responsáveis de concessão de crédito. Em resumo, este artigo destaca a importância crescente da aplicação de modelos preditivos e análises estatísticas no mercado de crédito para pessoa física, delineando um panorama integral para abordar desafios cruciais e promover a resiliência e eficiência nesse setor vital da economia.

2 OBJETIVOS

O objetivo deste estudo é desenvolver um modelo de aprendizado de máquina de avaliar o risco de crédito para pessoas físicas, aprimorando o processo de concessão de crédito no setor financeiro. Este modelo visa proporcionar uma abordagem mais precisa e eficiente para a tomada de decisões, avaliando o risco de inadimplência e otimizando a alocação de recursos financeiros.

3 JUSTIFICATIVA

A concessão de crédito é uma atividade crítica no setor financeiro, que exige uma análise minuciosa e precisa para equilibrar o risco financeiro e a satisfação do cliente. Com a crescente disponibilidade de dados e avanços em algoritmos de Machine Learning, é possível melhorar a precisão e eficiência na tomada de decisões de crédito. Este artigo visa contribuir para a otimização das operações de crédito, mitigando riscos financeiros e promovendo uma gestão mais eficaz de recursos, ao mesmo tempo em que atende às demandas de um mercado financeiro em constante evolução.

4 REVISÃO DA LITERATURA

4.1 RISCO E GERENCIAMENTO DE CRÉDITO

No mercado financeiro, um dos riscos mais proeminentes que as instituições financeiras enfrentam é o risco de crédito. O risco de crédito é a probabilidade de que um tomador de crédito não cumpra suas obrigações de pagamento, o que pode resultar em perdas financeiras para o credor. Esse risco é inerente a todas as transações de empréstimo e é crucial compreender e gerenciar adequadamente. (FIGUEREDO, 2001)

O risco de crédito não é uniforme, conforme explicado por Securato (2012) a incerteza quanto ao resultado é que cria a condição de risco na operação de conceder crédito. Esse risco abrange várias dimensões. E dentre os vários subtipos desse risco, um dos mais evidentes é o risco de inadimplência, que ocorre quando o tomador de crédito não consegue pagar o empréstimo na data acordada. Isso pode levar a perdas substanciais para o credor. (FIGUEREDO, 2001)

Além disso, Figueredo (2001) acrescenta o risco de atraso no pagamento, onde o devedor efetua pagamentos após a data de vencimento, resultando em custos adicionais para a instituição financeira. Por fim, há o risco de perda, que considera o valor que pode ser perdido em caso de insolvência.

Para Santos (2003), essa avaliação do risco de um potencial cliente pode ser feita por meio de julgamento, como uma forma mais subjetiva que envolve uma análise mais qualitativa. Ou por meio da classificação do tomador, via modelos de avaliação, envolvendo uma análise mais quantitativa.

Silva (2003) salienta a ideia de que gerenciamento de risco de crédito é o processo de avaliar, mitigar e monitorar os riscos associados ao empréstimo de dinheiro a indivíduos ou empresas. As abordagens tradicionais envolvem a avaliação minuciosa do histórico de crédito do tomador de crédito, análise de crédito detalhada, uso de garantias e políticas de concessão de crédito conservadoras ou liberais

Os autores Blatt (1999) e Securato (2012) complementam a ideia que gerenciamento de crédito voltado a pessoa física começa com a avaliação do perfil do solicitante. Isso inclui a análise do histórico de crédito, que fornece informações sobre o comportamento de pagamento passado do indivíduo. também consideram a pontuação de crédito do indivíduo, que é uma representação numérica da probabilidade de o indivíduo pagar suas dívidas.

Além disso, para Silva (2003) o gerenciamento de crédito envolve a avaliação da capacidade do indivíduo de pagar o empréstimo. Isso é feito analisando-se a renda do indivíduo, suas despesas e quaisquer outras obrigações financeiras que possam afetar sua capacidade de reembolsar o empréstimo.

Conforme evidenciado por Trapp & Corrar (2005) se um indivíduo começar a atrasar os pagamentos, a instituição financeira pode entrar em contato com ele para discutir as opções disponíveis para resolver a situação, mitigando ou até impedindo a entrada de um cliente para a inadimplência e conseqüentemente a insolvência. O autor Silva (2003) explica a importância do gerenciamento contínuo do crédito e define como essencial. Isso envolve o monitoramento regular do comportamento de pagamento do indivíduo e a intervenção precoce se surgirem problemas.

Ainda para o autor, a ênfase na importância da análise tanto qualitativa quanto quantitativa, considerando não apenas o histórico de crédito, mas também a capacidade financeira do indivíduo, acrescenta uma camada de profundidade ao entendimento desse cenário, destacando a necessidade de uma abordagem holística e proativa para mitigar as incertezas inerentes a essa atividade financeira crucial.

4.2 PRINCIPAIS DESAFIOS DO MERCADO DE CRÉDITO

O mercado de crédito é um pilar fundamental do setor financeiro, impulsionando o crescimento econômico ao fornecer acesso a financiamento para indivíduos e empresas, conforme a interpretado por Haussen (2022). No entanto, o mercado de crédito enfrenta uma série de desafios significativos que afetam sua estabilidade e eficácia. Entre os principais desafios, Brigham, Gapenski e Ehrhardt (2001) destacam-se a inadimplência seguida pela Insolvência, as fraudes e a regulamentação.

Inadimplência: é um termo que ecoa profundamente no mercado de crédito e pode ser definida, de acordo com Securato (2012), como a falha de um devedor em cumprir com suas obrigações de pagamento em um empréstimo, seja ele um financiamento para a compra de uma casa, um empréstimo pessoal, ou um cartão de crédito. Essa situação não é apenas um desafio financeiro, mas também uma preocupação crítica para os credores e para o funcionamento eficaz do mercado de crédito como um todo.

Os impactos da inadimplência são vastos e variados. Em primeiro lugar, Silva (2003) evidencia as perdas financeiras diretas para o credor. Quando um devedor deixa de efetuar pagamentos, o valor do empréstimo não é reembolsado conforme o previsto, causando uma

diminuição direta nos ativos do credor. Isso pode ser particularmente prejudicial se a inadimplência ocorrer em grande escala ou envolver empréstimos significativos.

Além disso, a inadimplência gera custos adicionais para o credor, o processo de recuperação de ativos de inadimplentes envolve despesas legais, administrativas e de cobrança, todas as quais diminuem ainda mais os ganhos potenciais do credor. A inadimplência também afeta a rentabilidade da instituição financeira, uma vez que o lucro líquido é diretamente afetado pelas perdas associadas à falta de pagamento (TRAPP & CORRAR (2005)).

Silva (2003) mostra outro impacto da inadimplência é o impacto negativo sobre a saúde financeira das instituições financeiras. Quando ocorrem inadimplências em grande escala, o capital disponível para empréstimos pode ser reduzido, uma vez que as perdas podem exceder as reservas destinadas a cobrir essas perdas. Isso, por sua vez, pode restringir a capacidade do credor de fornecer financiamento a outros tomadores de crédito e, conseqüentemente, afetar o crescimento econômico.

Além das implicações financeiras, a inadimplência também prejudica a reputação do credor. Os devedores podem divulgar informações negativas sobre a instituição, o que pode afastar futuros tomadores de crédito afetando negativamente a confiança do público na instituição financeira. (SILVA, 2003)

Insolvência: é uma condição em que um devedor se torna incapaz de cumprir suas obrigações financeiras, resultando na incapacidade de pagar suas dívidas no prazo acordado. Esta situação é um dos maiores riscos enfrentados pelo mercado de crédito, pois pode ter impactos substanciais nas instituições financeiras e no sistema como um todo. (SECURATO 2012)

Para Cherobim (2011), a insolvência pode ser desencadeada por uma variedade de fatores, incluindo a perda de emprego que pode tornar difícil ou impossível para um devedor cumprir suas obrigações de pagamento. Eventos Inesperados como desastres naturais, problemas de saúde súbitos ou outras emergências podem levar a despesas imprevistas, dificultando o pagamento de dívidas. Alterações nas condições econômicas globais ou locais podem impactar a capacidade dos devedores de pagar suas dívidas

Pela interpretação de Ferraz (2023), uma das principais causas subjacentes de insolvência no contexto do mercado é a má gestão financeira. Este fenômeno resultado de decisões inadequadas e falta de prudência nas práticas financeiras individuais, contribuem significativamente para o desafio enfrentado pelas instituições financeiras.

A incapacidade de controlar os gastos é uma característica comum da má gestão financeira. Onde tomadores de crédito que vivem além de suas possibilidades estão mais propensos a entrar em situações de insolvência devido ao desequilíbrio entre receitas e despesas. (FERRAZ, 2023)

Ainda para o Autor, Tomar empréstimos sem avaliar adequadamente a capacidade de pagamento é uma forma de má gestão financeira assim como decisões de investimentos imprudentes em ativos de alto risco sem avaliar adequadamente os potenciais retornos e riscos associados. Podendo levar a perdas significativas, contribuindo para a insolvência, O endividamento excessivo pode levar a um fardo financeiro insustentável.

A ausência de uma reserva financeira para enfrentar emergências é uma marca registrada da má gestão financeira. Quando ocorrem eventos imprevistos, como desemprego ou despesas médicas inesperadas, os devedores podem se ver incapazes de cumprir suas obrigações financeiras. (CHEROBIM (2011)

Segundo Cherobim (2011), esses hábitos, dificultam ou até impossibilitam a capacidade do devedor de cumprir suas obrigações aumentando o endividamento e obrigando-o a priorizar as dívidas mais cruciais o que pode levar ao não pagamento de obrigações, como hipotecas, pagamentos de cartões de crédito e empréstimos.

Alguns tomadores de crédito vão à procura de empréstimos adicionais com o objetivo de cobrir déficits financeiros, criando um ciclo vicioso, resultando em atrasos, falta de pagamentos podendo chegar até na possibilidade de vendas forçadas de ativos para cobrir dívidas. Isso pode incluir a venda de propriedades ou outros ativos para liquidar passivos. (TRAPP & CORRAR, 2005)

Conforme evidenciado por Trapp & Corrar (2005) e Silva (2003), A má gestão financeira aumenta os custos de recuperação para os credores, resultando em resistência dos devedores durante os processos de cobrança. Isso acarreta significativos custos para as instituições financeiras impactando sua rentabilidade. Com a rentabilidade comprometida, a impossibilidade de recuperação total dos fundos emprestados prejudica a geração de lucros, afetando diretamente os resultados financeiros do credor.

Para Silva (2003), em cenário onde as perdas atingem uma grande escala, as instituições financeiras podem enfrentar uma restrição de recursos disponíveis para empréstimos futuros. A confiança na recuperação diminuída pode levar a uma abordagem mais cautelosa na concessão de crédito.

Ainda segundo o autor, o gerenciamento eficaz do risco relacionado à má gestão financeira envolve a promoção de educação financeira, a implementação de práticas de concessão de crédito responsáveis e a incorporação de modelos analíticos avançados para avaliação do risco individual. Instituições financeiras podem adotar estratégias proativas para orientar tomadores de crédito em práticas financeiras saudáveis, ajudando a prevenir a insolvência decorrente da má gestão financeira.

Outra causa com o impacto significativo na insolvência de um devedor, são taxas de juros elevadas desempenhando um papel significativo na determinação da saúde financeira dele. O impacto direto dessas taxas sobre os custos de empréstimos pode ser substancial, aumentando os riscos de inadimplência e insolvência. (CAMARGO, 2007)

Conforme interpretado por Camargo (2007) e Cherobim (2011), taxas de juros elevadas resultam em encargos financeiros maiores para os devedores. À medida que esses encargos aumentam, a capacidade do devedor de cumprir suas obrigações financeiras pode ser comprometida. Essas taxas altas tornam o crédito menos acessível para muitos tomadores de crédito, especialmente aqueles com histórico de crédito menos robusto. Isso pode levar a uma dependência de fontes mais caras de financiamento que contribuem em conjunto, a criação de um ciclo de endividamento, à medida que os devedores, ao tentar gerenciar suas dívidas existentes, recorrem a novos empréstimos para cobrir pagamentos e custos diários, aumentando ainda mais sua carga financeira.

Em resumo, a insolvência impacta o credor com perdas financeiras diretas e efeitos indiretos, como ameaça à reputação. Estratégias eficazes de recuperação, concessão de crédito responsável e preservação da confiança pública são essenciais. A abordagem proativa inclui educação financeira aos tomadores de crédito para evitar má gestão. Políticas regulatórias, como limites de taxas de juros, podem proteger e contribuir para a estabilidade do mercado de crédito.

Fraudes: para Securato (2003), as fraudes representam um desafio substancial no mercado de crédito, criando um ambiente de incerteza e potencialmente prejudicando tanto credores quanto tomadores de crédito. A fraude ocorre quando um indivíduo, com má-fé, fornece informações falsas, enganosas ou fraudulentas ao solicitar um empréstimo ou ao usar um cartão de crédito. Esse tipo de atividade enganosa pode ter sérios impactos financeiros e reputacionais.

Para os credores, a fraude resulta em perdas financeiras diretas. Quando as instituições concedem crédito com base em informações falsas, elas correm o risco de não recuperar o valor

emprestado. Além disso, a detecção e a investigação de fraudes consomem recursos significativos, incluindo pessoal e tecnologia, aumentando os custos operacionais. (Ferraz, 2023)

Ainda para o autor, para os tomadores de crédito legítimos, a fraude pode ter impactos prejudiciais. A identificação incorreta em uma fraude de crédito pode resultar em uma pontuação de crédito mais baixa, dificultando a obtenção de crédito no futuro. Além disso, as vítimas de fraude muitas vezes enfrentam um processo burocrático demorado para limpar seus registros de crédito e resolver disputas.

Regulamentação: Segundo Veríssimo (2023), desempenha um papel crítico no mercado de crédito. Reguladores financeiros estabelecem diretrizes e requisitos que as instituições devem cumprir para garantir práticas justas e seguras de concessão de crédito. O não cumprimento das regulamentações pode resultar em sanções financeiras substanciais e danos à reputação das instituições.

Tendo visto os desafios a serem enfrentados, Provost & Fawsett (2016) afirmam que a ciência de dados e a análise estatística podem fornecer insights valiosos para a criação de estratégias de gerenciamento mais eficazes. Através da coleta e análise de dados em tempo real, as instituições podem tomar decisões de crédito mais informadas, otimizar a alocação de recursos e aprimorar a prevenção de fraudes.

Em suma, os principais desafios do mercado de crédito, como inadimplência, fraudes e regulamentação, são questões críticas para as instituições financeiras que possuem uma complexidade elevada no entendimento de suas causas. Nesse contexto, a ciência de dados e a estatística apresentam ótimas técnicas para a abordagem precisa no entendimento desses desafios, fornecendo ferramentas e técnicas avançadas para mitigar riscos, melhorar a eficiência e garantir a conformidade regulatória.

4.3 CIÊNCIA DE DADOS: ANALÍTICA, PREDITIVA E PRESCRITIVA.

A ciência de dados é uma área interdisciplinar que se desdobra em diversas abordagens, cada uma desempenhando um papel fundamental na interpretação e utilização eficaz dos dados e tem crescido significativamente nas últimas décadas. (PROVOST & FAWSETT, 2016)

O crescimento da internet, das redes sociais e dos dispositivos móveis está gerando uma quantidade cada vez maior de dados, criando oportunidades para a ciência de dados em

uma ampla gama de aplicações que vão desde previsão de tendências como perfil de clientes, abrindo oportunidades de crescimento, até a automatização de processos e escolha dos melhores caminhos reduzindo custos. (GRUS, 2021)

Ainda segundo o Autor, esse crescimento tem beneficiado muitas áreas, em especial, instituições financeiras, que encontraram uma forma de extrair insights e conhecimento de dados. Esses insights podem ser usados para melhorar a tomada de decisões, desenvolver novos produtos e serviços, ou simplesmente entender melhor o mundo ao nosso redor.

Provost & Fawsett (2016) traz argumentos defendendo que a ciência de dados é uma área constante evolução, pois à medida que novas técnicas e tecnologias são desenvolvidas. Isso abre novas possibilidades para a área, e é provável que a ciência de dados continue a desempenhar um papel cada vez mais importante em tomadas de decisões do mercado financeiro.

A ciência de dados, desdobra-se em várias perspectivas, cada qual desempenhando um papel crucial na interpretação e aplicação eficaz dos dados. Essas perspectivas envolvem a Ciência Analítica, a Ciência Preditiva e a Ciência Prescritiva.

Ciência Analítica: constitui a base da ciência de dados, focando na análise retrospectiva de dados histórico. É uma abordagem essencial que se dedica à análise minuciosa de conjuntos de dados para extrair insights valiosos. Essa vertente se destaca por seu foco na compreensão detalhada dos dados, utilizando métodos estatísticos, técnicas de mineração de dados e algoritmos avançados. O objetivo principal da Ciência Analítica é revelar padrões, tendências e correlações nos dados, proporcionando uma visão aprofundada do cenário analisado. (MCKINNEY, 2023)

Ainda segundo o autor, conforme explicado em sua obra, a ciência Analítica inicia-se muitas vezes com uma análise descritiva, que busca resumir e descrever características fundamentais do conjunto de dados, como médias, medianas e desvios-padrão. Esse processo proporciona uma visão geral, permitindo uma compreensão inicial dos dados.

Segundo Grus (2021) Após entender as características do conjunto de dados com técnicas de análise descritivas, é iniciada a análise exploratória de dados, quem utilizam de técnicas gráficas e estatísticas para investigar padrões e relacionamentos nos dados. Gráficos, box plots e correlações são ferramentas comuns nesse estágio, oferecendo uma compreensão mais aprofundada das relações entre variáveis.

Para validar hipóteses e inferir conclusões sobre a população, a Ciência Analítica emprega diversos testes estatísticos. Tais testes são cruciais para garantir que as descobertas são estatisticamente significativas e não resultado do acaso (MUELLER, 2021).

A Ciência Analítica desempenha um papel fundamental no mercado financeiro, capacitando as instituições conhecerem seus dados com precisão e eficácia. Seja na gestão de riscos, detecção de fraudes ou na personalização de serviços, essa abordagem analítica colabora para decisões fundamentadas em dados estatísticos e estratégias mais sólidas. (MCKINNEY, 2023)

Ciência Preditiva: esta vai além da análise retrospectiva, empregando algoritmos avançados para prever tendências futuras e comportamentos com base em padrões identificados em conjuntos de dados históricos, essa abordagem busca antecipar cenários e comportamentos futuros, proporcionando insights valiosos para a tomada de decisões. (MCKINNEY, 2023)

Os autores Grus (2021), Provost & Fawcett (2016) e Géron (2019) destacam algumas ferramentas como as mais fundamentais da ciência preditiva, sendo elas:

Modelos Preditivos: A construção de modelos é o cerne da Ciência Preditiva. Algoritmos como regressão linear, árvores de decisão, redes neurais e máquinas de suporte vetorial são aplicados para criar modelos que podem prever resultados futuros com base em variáveis de entrada.

Análise de Séries Temporais: Em muitos contextos, especialmente no financeiro e no climático, a análise de séries temporais é fundamental. Ela examina dados sequenciais ao longo do tempo para identificar padrões sazonais, tendências e variações cíclicas, permitindo previsões precisas.

Mineração de Dados: A Ciência Preditiva emprega técnicas avançadas de mineração de dados para descobrir padrões ocultos em grandes conjuntos de dados. Isso inclui a identificação de correlações complexas que podem ser utilizadas para prever eventos futuros.

Aprendizado de Máquina: Algoritmos de aprendizado de máquina desempenham um papel crucial na Ciência Preditiva, permitindo que sistemas automatizados melhorem seu desempenho ao longo do tempo, ajustando-se a novos dados e refinando suas previsões.

Em síntese, a ciência preditiva oferece uma visão futurista para o setor financeiro, capacitando as instituições a anteciparem cenários, tomarem melhores decisões e se adaptarem rapidamente às mudanças no ambiente econômico, seja na gestão de investimentos, avaliação de riscos ou personalização de serviços, essa abordagem preditiva é uma ferramenta indispensável para enfrentar os desafios dinâmicos do mundo financeiro.

Ciência Prescritiva: fornece recomendações acionáveis para otimizar decisões e estratégias enquanto a ciência analítica interpreta dados passados e a Ciência Preditiva antecipa o que pode acontecer, a ciência prescritiva recomenda ações específicas para otimizar resultados, considerando uma variedade de variáveis e restrições (GRUS, 2021).

Ao contrário de apenas fornecer insights ou previsões, a Ciência Prescritiva busca otimizar decisões complexas, para tanto utiliza-se de algoritmos avançados para recomendar ações específicas que levarão a resultados desejados, considerando diferentes cenários e objetivos (GÉRON, 2019).

Ainda para o autor, a capacidade de modelar diferentes cenários é uma característica distintiva da ciência prescritiva, permitindo que as organizações avaliem o impacto de diferentes decisões sob diversas condições como restrições orçamentárias, limitações de recursos e políticas específicas, garantindo que as recomendações se alinhem com a viabilidade prática, proporcionando uma compreensão abrangente das implicações de suas escolhas.

Algumas ferramentas da ciência prescritiva destacada pelos autores Grus (2021) e Géron (2019) são:

Programação Linear e Não Linear: Estas técnicas são fundamentais na Ciência Prescritiva para resolver problemas complexos de otimização, onde é necessário maximizar ou minimizar uma função objetivo sujeita a várias restrições.

Algoritmos Genéticos e Redes Neurais: Inspirados no processo de evolução biológica, os algoritmos genéticos são aplicados na otimização de soluções em uma ampla gama de problemas, adaptando-se ao longo do tempo para encontrar as melhores respostas, enquanto as redes neurais, além de terem as mesmas características de um algoritmo genético, podem recomendar ações específicas com base em padrões identificados nos dados

A Ciência Prescritiva, ao oferecer recomendações acionáveis e otimizadas, eleva a ciência de dados a um patamar onde não apenas se compreende e antecipa, mas também se influencia proativamente o curso dos eventos, suas aplicações abrangentes a torna uma ferramenta valiosa para aprimorar a eficiência e a eficácia em uma variedade de setores, moldando o futuro das tomadas de decisões estratégicas. (GÉRON, 2019).

Por fim, integrar esses campos de conhecimento da ciência de dados nas instituições financeiras, pode-se construir abordagens mais holísticas para a concessão de crédito, onde análise retrospectiva fornece um entendimento sólido do passado, a predição antecipa desafios futuros e a prescrição oferece diretrizes práticas para uma gestão de riscos mais eficiente. A convergência de disciplinas analíticas não apenas fortalece a capacidade das instituições

financeiras de lidar com os desafios do mercado de crédito, mas também contribui para a construção de um sistema financeiro mais resiliente e sustentável.

4.4 O USO DA CIÊNCIA DE DADOS NA TOMADA DE DECISÕES.

A ciência de dados tem revolucionado a maneira como as organizações tomam decisões. Ao utilizar técnicas avançadas de análise e previsão de dados, as empresas podem obter insights valiosos que podem ajudá-las a tomar melhores decisões e mais eficazes. Para os autores Provost & Fawsett (2016), também Fávero (2017), a ciência de dados pode ser dividida por campos de compreensão como:

Compreensão do cliente: Na compreensão dos clientes, por exemplo, ela permite que as empresas acessem insights detalhados sobre o comportamento, as preferências e as necessidades dos clientes. Essas informações podem ser usadas para personalizar produtos e serviços, melhorar a experiência do cliente e aumentar a lealdade à marca, e no caso do crédito, pode ser usada para moldar novas políticas de regulamentação de crédito

Avaliar o comportamento do cliente é importante para o mercado de concessão de crédito, onde a ciência de dados se concentra em entender como os clientes interagem com a instituição financeira. Isso pode incluir tudo, desde a forma como o cliente toma o crédito até o comportamento de pagamentos, bem como a entrada na inadimplência, já que analisar esses comportamentos, as instituições financeiras podem identificar padrões e tendências que podem mudar a forma como é visualizado o risco para conceder crédito. (PROVOST & FAWSETT, 2016)

A segmentação de clientes é outra área onde a ciência de dados pode fornecer insights valiosos, ao agrupar os clientes com base em características semelhantes, as instituições financeiras podem criar estratégias mais direcionadas e personalizadas. Isso não só melhora a eficácia do gerenciamento de crédito, mas também pode levar a uma maior satisfação do cliente. (PROVOST & FAWSETT, 2016)

Para Fávero (2017), A ciência de dados também permite um alto grau de personalização. Com base nos dados coletados sobre o comportamento e as preferências dos clientes, as instituições financeiras podem personalizar suas ofertas para atender às necessidades específicas de cada cliente.

Isso pode aumentar a relevância das ofertas para o cliente e, por sua vez, aumentar as taxas de conversão identificar clientes que possam estar em risco de se tornarem inadimplente

ou solicitações de crédito fraudulentas preservando a segurança e a saúde financeira. já que dependendo do seguimento, fraudadores podem assumir um comportamento semelhante. Essas previsões podem permitir que as empresas tomem medidas proativas para engajar seus clientes e atender às suas necessidades como também preventivas, fugindo do risco de perdas financeiras. (PROVOST & FAWSETT, 2016)

Otimização de Operações: Para Fávero (2017), com os insights extraídos, a empresa pode tomar medidas para otimizar esses processos, usando modelos analíticos a ciência de dados pode ajudar as empresas a identificarem ineficiências em suas operações e a tomar medidas para melhorá-las, onde ao analisar os dados operacionais, uma financeira pode descobrir que certa regra dentro de uma política pode ser explorada por fraudadores.

Ainda para o autor, outra contribuição está na redução de, ao analisar os dados de custos as empresas podem identificar áreas onde estão gastando mais do que o necessário e buscar maneiras de reduzir esses custos, isso pode envolver coisas como renegociar contratos com fornecedores, investir em tecnologia para automatizar tarefas caras ou reestruturar as operações para eliminar desperdícios.

Por meio de informações, ao invés do gestor ou analista confiar na intuição ou na experiência passada, eles podem usar os insights derivados dos dados para tomar decisões baseadas em evidências, levando a melhores decisões e criando uma cultura de responsabilidade e transparência, o que possibilita na autonomia da máquina para tomada de decisões baseadas em dados. (PROVOST & FAWSETT, 2016)

Mitigação de Riscos: Conforme os autores Provost & Fawsett (2016) e Fávero (2017), as técnicas de ciência de dados pode ajudar as empresas a identificarem riscos potenciais que podem não ser imediatamente aparentes, por exemplo, ao analisar grandes volumes de dados, uma empresa pode identificar tendências ou padrões que indicam um risco emergente. Isso pode permitir que a empresa tome medidas para mitigar esse risco antes que ele se torne um problema.

Ainda para os autores, a ciência de dados pode ajudar nesta avaliação ao fornecer ferramentas e técnicas para analisar e quantificar o risco, uma vez identificados os riscos, a ciência de dados pode ajudar na sua avaliação. Isso envolve a determinação da probabilidade de o risco se concretizar e do impacto potencial que ele poderia ter sobre a empresa.

Após o risco ser mensurado, a instituição tem a capacidade de gerenciar políticas para lidar com o risco, abarcando o desenvolvimento e a implementação de estratégias para lidar com os riscos identificados. Essas estratégias podem incluir a transferência do risco para outra

parte (por exemplo, através de seguros), a redução do risco (por exemplo, através da implementação de controles internos) ou a aceitação do risco. PROVOST & FAWSETT (2016)

Por fim, a ciência dos dados pode ajudar no monitoramento contínuo dos riscos e das decisões. Isso envolve o acompanhamento regular e a avaliação da eficácia das estratégias de gerenciamento de riscos implementadas, de maneira que ela pode facilitar esse monitoramento ao fornecer ferramentas para coletar, analisar e visualizar dados sobre os riscos e decisões.

Em resumo, a ciência de dados oferece ferramentas poderosas para impulsionar a produtividade empresarial, utilizando algoritmos para otimizar programações de trabalho, empregando machine learning em previsões de comportamento do cliente e antecipando eventos econômicos. Essas técnicas possibilitam insights valiosos, aprimorando a compreensão do cliente, otimizando operações e mitigando riscos, permitindo decisões mais informadas e eficazes para alcançar melhores resultados.

4.5 CIÊNCIA DE DADOS E SEU PAPEL NA TOMADA DE DECISÕES ESTRATÉGICAS.

Com uma gama de ferramentas a disposição. Este estudo explorou algumas dessas técnicas e modelos, destacando suas fundamentações teóricas e práticas pelos autores McKinley (2023), Mueller (2016), Provost & Fawsett (2016), Fávero (2017),

Análise Descritiva: Para McKinley (2023), a fase inicial da exploração de dados, é essencial para obter uma compreensão aprofundada das características fundamentais presentes no conjunto de dados, onde são empregadas diversas técnicas, incluindo estatísticas descritivas como médias, medianas e desvios-padrão, juntamente com visualizações gráficas, como histogramas e box plots.

Essas técnicas estatísticas não apenas resumem os dados de maneira precisa, mas também proporcionam uma representação visual que facilita a identificação de padrões preliminares. Dessa forma, a Análise Descritiva não apenas resume, mas também ilustra de forma elucidativa, preparando o terreno para análises mais aprofundadas ao longo do processo de exploração de dados.

Análise Exploratória de Dados: Ainda com o raciocínio do autor, trata-se de uma etapa subsequente que vai além da simples descrição, aprofundando a compreensão do conjunto de dados por meio de técnicas visuais e estatísticas. Nessa fase, são empregados gráficos, box plots e análises de correlação para identificar padrões, outliers e relações entre variáveis.

Mueller (2016), explica que as técnicas visuais e estatísticas citadas são fundamentais para proporcionar insights mais detalhados, permitindo uma análise mais aprofundada e direcionando a escolha de abordagens analíticas mais avançadas desempenhando um papel crucial na identificação de nuances e complexidades que podem guiar investigações mais detalhadas ao longo do processo analítico.

Engenharia de dados: representa um estágio mais abrangente na manipulação e preparação de dados, indo além do pré-processamento básico. Esta fase incorpora a transformação significativa dos dados, incluindo a criação de novas variáveis, a agregação de dados e a seleção criteriosa de características relevantes para o problema em análise. (GRUS, 2021)

Ainda do ponto de vista teórico do autor, a engenharia de dados abrange conceitos avançados de otimização e estruturação eficiente de conjuntos de dados, implicando não apenas na resolução de problemas específicos relacionados à qualidade dos dados, mas também na melhoria da eficiência computacional durante análises posteriores.

A seleção de características relevantes requer uma compreensão profunda do problema em questão e a identificação das variáveis mais impactantes, sendo necessários em alguns casos a criação de novas variáveis envolvendo a combinação ou transformação de variáveis existentes para extrair informações mais úteis ou representativas. (MUELLER, 2016).

Para Géron (2019) a engenharia de dados também está relacionada à otimização do desempenho computacional, especialmente quando lidamos com conjuntos de dados, estratégias como a escolha eficiente de estruturas de dados e algoritmos podem contribuir significativamente para acelerar processos analíticos subsequentes.

Assim, a engenharia de dados não apenas aprimora a qualidade dos dados, mas também os molda de maneira a otimizar a eficiência e a relevância para os objetivos específicos da análise.

Pré-processamento de dados: Esta etapa desempenha um papel crucial na preparação dos dados para análises avançadas, visando garantir a qualidade e confiabilidade das conclusões extraídas que abrange uma variedade de tarefas, desde lidar com valores ausentes até normalizar escalas, converter dados categóricos em formatos apropriados e tratar outliers. (GÉRON, 2019).

Do ponto de vista do autor, o pré-processamento de dados requer uma compreensão aprofundada de técnicas estatísticas e computacionais. Isso inclui o conhecimento de métodos estatísticos para imputação de valores ausentes, algoritmos de normalização e transformação de

dados, além de estratégias para lidar com valores atípicos, a prática, por outro lado, envolve a aplicação efetiva dessas técnicas utilizando ferramentas específicas de processamento de dados.

Ao lidar com valores ausentes, por exemplo, estratégias como imputação baseada em média, mediana ou algoritmos mais avançados podem ser empregadas. A normalização de escalas é crucial para garantir que variáveis com unidades diferentes não influenciem desproporcionalmente as análises (MUELLER, 2016).

Regressão Logística: Sendo mais específico em relação às ferramentas, o modelo computacional de Regressão Logística é um método estatístico amplamente utilizado em Aprendizado de Máquina, ele utiliza a função logística para modelar a probabilidade de uma determinada classe ou evento ocorrer, esta abordagem é particularmente valiosa em problemas de classificação binária, onde o resultado desejado é de natureza dicotômica ou binária. (GÉRON, 2019)

Matematicamente, conforme explicado por Pelegrina (2023), a função logística, também conhecida como curva S, é essencial para a Regressão Logística, pois permite transformar uma combinação linear de variáveis independentes em uma probabilidade condicional. Isso é expresso pela fórmula:

$$p(x) = \frac{1}{1 + e^{-z}}$$

Onde e é a base do logaritmo natural, número de Euler que equivale a aproximadamente 2,7183. E z é o resultado da função linear dada por:

$$z = \beta_0 + \sum \beta_i \theta_i$$

Sendo $p(x)$ a probabilidade do objeto x de acontecer.

Alguns pontos positivos de seu uso, incluem interoperabilidade, pois os coeficientes fornecem uma interpretação direta da relação entre variáveis independentes e a variável dependente, eficiência computacional se comparada a modelos mais complexos, a Regressão Logística é relativamente simples e eficiente (PELEGRINA, 2023).

Ainda segundo a autora, é um modelo suscetível a outliers ou valores atípicos nos dados, o que exige que as etapas de pré-processamento, sejam cumpridas com eficácia. Ele possui algumas limitações de uso quanto a problemas de classificação pois ele entre as variáveis independentes e os pesos (log-odds) da variável dependente

Essa ferramenta frequentemente utilizada para problemas de classificação binária, como prever se um cliente comprará ou não um produto.

Random Forest: é uma técnica de conjunto que constrói múltiplas árvores de decisão e as combina para obter uma predição mais robusta. Cada árvore é construída de maneira independente treinada em uma amostra aleatória dos dados, para fazer previsões as árvores individuais votam (no caso de classificação) ou contribuem com médias (no caso de regressão). A previsão final é determinada pela maioria das votações (ou média) entre as árvores. (GÉRON, 2019)

O processo de decisão durante a construção de cada árvore de decisão é resultado de uma função que calcula o ganho de informação, na qual em cada nó da árvore, o algoritmo escolhe a divisão que possui a menor desordem nos dados, ou seja, aquela que organiza melhor os dados (BEHNIA, 2023)

A função de ganho é dada por:

$$ganho = entropia_{(y)} - \sum peso_{\beta_n} + entropia_{\beta_n}$$

É amplamente utilizado em problemas de classificação e regressão. Sua capacidade de lidar com dados não lineares e identificar interações complexas torna-o valioso para uma variedade de aplicações (BEHNIA, 2023).

Entre suas vantagens estão a robustez, capacidade de lidar com grande quantidade de variáveis e resistência ao overfitting. Suas fraquezas incluem a dificuldade de interpretação em comparação com modelos mais simples e maior consumo computacional (GÉRON, 2019)

Extreme Gradient Boost: é uma implementação do algoritmo Gradient Boosting que utiliza árvores de decisão como base. Ele cria uma sequência de modelos, onde cada modelo subsequente corrige os erros do modelo anterior, focando nas instâncias que foram classificadas incorretamente (SIKANDER, 2022).

Ainda segundo o autor, Extreme Gradient Boost é amplamente utilizado em situações em que a precisão é crucial. Pode ser aplicado tanto a problemas de classificação quanto de regressão. Sendo assim suas vantagens em uso é a alta precisão, eficiência e capacidade de lidar com dados ausentes. Contras envolvem o risco de overfitting e a necessidade de ajuste cuidadoso dos hiper parâmetros.

Técnicas de Avaliação de Modelos: Diversas técnicas são empregadas para avaliar a eficácia dos modelos, incluindo a validação cruzada, a matriz de confusão, a curva ROC e métricas como precisão, recall, F1-score e área sob a curva (AUC). Cada uma dessas técnicas fornece uma perspectiva única sobre o desempenho do modelo em diferentes cenários (GÉRON, 2019).

Essas técnicas e modelos capacitam a ciência de dados a desempenhar um papel central na tomada de decisões estratégicas, permitindo que as organizações extraiam insights significativos e informação preditiva a partir de dados complexos, contribuindo para uma abordagem mais informada e eficaz na resolução de problemas e na definição de estratégias futuras (GÉRON, 2019).

5 DESENVOLVIMENTO

5.1 AQUISIÇÃO DE DADOS

A obtenção de dados é uma etapa crucial em qualquer projeto de ciência de dados. Neste estudo, os dados foram originados a partir de um conjunto disponível no Kaggle, uma plataforma amplamente reconhecida por hospedar conjuntos de dados diversificados para estudos e competições de aprendizado de máquina. A fonte específica utilizada foi o conjunto de dados "Credit Card" disponível em (<https://www.kaggle.com/datasets/mishra5001/credit-card>).

O formato do conjunto de dados é em CSV (Comma-Separated Values), uma escolha comum para a representação tabular de dados. No caso em questão, o conjunto contém duas tabelas, embora, para este estudo, tenha sido focado principalmente em uma delas.

O volume total de dados é considerável, abrangendo mais de 400 mil linhas distintas. Essa magnitude coloca o conjunto de dados na categoria de grandes conjuntos, o que proporciona desafios e oportunidades distintas durante o processo de análise.

O conjunto de dados, proveniente de 2019, oferece uma visão histórica, embora não tenha sido identificada nenhuma sazonalidade ou eventos específicos relevantes. A presença de mais de 60 variáveis categóricas, detalhando o comportamento de clientes em relação a empréstimos, adiciona complexidade e riqueza à análise.

Uma descrição detalhada das variáveis presentes no conjunto de dados está disponível, fornecendo insights sobre a natureza e o propósito de cada atributo. Além disso, o conjunto possui algumas peculiaridades, como notas (scoring) de região de moradia e colunas com valores lógicos (1 ou 0), acrescentando uma camada de detalhes à sua estrutura.

Os dados são de domínio público, disponíveis para uso gratuito, sem restrições específicas associadas. Apesar de não haver atualizações regulares, o conjunto oferece um instantâneo relevante para análises retrospectivas.

A aquisição desses dados envolveu a escolha de uma fonte confiável, a compreensão da estrutura do conjunto, a identificação de peculiaridades e a confirmação de que atendem aos requisitos específicos do projeto de ciência de dados. Essa etapa estabelece a base para as análises subsequentes, proporcionando um entendimento sólido do contexto em que as decisões estratégicas serão fundamentadas.

5.2 PRÉ-PROCESSAMENTO DE DADOS

A fase de pré-processamento de dados desempenha um papel crucial na preparação adequada do conjunto de dados para análises mais avançadas, esta etapa é essencial para garantir a integridade e qualidade dos dados, bem como para permitir interpretações significativas nos estágios subsequentes da análise.

A seleção criteriosa das variáveis é fundamental para o sucesso do estudo, dentre as variáveis disponíveis, foram escolhidas aquelas que fornecem informações cruciais sobre os clientes, incluindo renda, histórico de crédito, emprego e características demográficas. Essa escolha foi norteadada pela relevância desses atributos na avaliação de solicitações de empréstimo e comportamento de pagamento.

A identificação e tratamento adequado de valores ausentes são considerações cruciais, para variáveis quantitativas, optou-se pela mediana em detrimento da média, dada a sensibilidade desta última a valores extremos. Já para as variáveis qualitativas, a técnica de codificação one-hot (`pd.get_dummies()`) foi empregada, assegurando uma integração eficaz dessas variáveis nas análises.

Variáveis categóricas foram transformadas por meio da criação de variáveis fictícias, uma técnica conhecida como one-hot encoding, onde essa abordagem permite que variáveis categóricas sejam incorporadas nas análises sem impor uma ordem artificial entre as categorias.

A normalização é um passo importante para garantir que variáveis em diferentes escalas contribuam igualmente para as análises, assim diversas técnicas foram testadas, e o `MinMaxScaler` foi adotado, preservando as características originais dos dados e ajustando os valores para um intervalo específico.

Técnica de Escalonamento	Fórmula Matemática	Explicação
MinMaxScaler	$X = \frac{X - X_{min}}{X_{max} - X_{min}}$	Reduz cada valor para um intervalo entre 0 e 1, com base nos valores mínimo e máximo originais. Útil quando se deseja manter a forma da distribuição original e os outliers são relevantes.

StandardScaler	$X \frac{X - \mu}{\sigma}$	Padroniza os dados para ter uma média (μ) de 0 e um desvio padrão (σ) de 1. Útil quando os dados seguem uma distribuição normal e não são sensíveis a outliers.
RobustScaler	$X = \frac{X - Q_1}{Q_3 - Q_1}$	Semelhante ao MinMaxScaler, mas utiliza o primeiro quartil (Q_1) e o terceiro quartil (Q_3) para torná-lo robusto a outliers. Útil quando os dados têm outliers significativos.
Normalizer	$X = \frac{X}{\sqrt{\sum \beta_n}}$	Reduz cada linha de dados para um comprimento de 1. Útil quando a magnitude dos dados não é importante, apenas a direção, sendo β_n cada amostra

Quadro 1 - Técnicas de Escalonamento

Fonte: elaborado pelo autor.

Apesar da ausência significativa de outliers, uma estratégia pré-definida consiste em substituí-los pela mediana, caso sejam identificados. Isso visa preservar a robustez das análises diante de valores extremos.

A engenharia de dados foi empregada na criação de 105 novas variáveis através da técnica de codificação one-hot. O que ampliou o conjunto de dados, proporcionando uma representação mais abrangente das variáveis categóricas originais.

5.3 CONSTRUÇÃO MANUAL DO MODELO DE REGRESSÃO LOGÍSTICA

A etapa de construção manual do modelo de Regressão Logística foi iniciada após a aquisição e pré-processamento dos dados. Nesta fase, buscou-se entender a dinâmica do modelo por meio de uma implementação direta, utilizando ferramentas Python e conceitos matemáticos fundamentais

Durante a construção do modelo, não foram estabelecidos critérios específicos para a escolha do modelo de Regressão Logística, pois o foco estava na compreensão dos conceitos e na implementação prática.

Essa implementação envolveu a criação de manual das funções fundamentais para a Regressão Logística com base na matemática função logística portanto foram utilizadas funções como ``insira_uns``, que adicionava uma coluna de uns aos dados para lidar com o termo de interceptação, ``sigmoid`` para aplicar a função sigmoideal aos resultados, ``função_custo`` para calcular a função de custo, ``gradient_descendent`` para realizar a otimização dos pesos e, finalmente, ``predicao`` para gerar previsões binárias.

Os parâmetros do modelo foram ajustados através de uma função que implementa o algoritmo de gradiente descendente, o processo foi iterativo, buscando minimizar a função de custo, mas não houve uma sintonia fina eficaz dos hiper parâmetros.

Os resultados iniciais do modelo construído manualmente foram inconsistentes para o conjunto de dados em questão, dado a complexidade e a diversidade do conjunto de dados demandavam uma abordagem mais sofisticada para obter resultados satisfatórios.

A construção manual do modelo destacou a importância de considerar cuidadosamente os hiper parâmetros e técnicas de otimização para obter um modelo preciso e robusto. Além disso, demonstrou a complexidade inerente à implementação manual de algoritmos, especialmente em conjuntos de dados desafiadores.

O desempenho geral do modelo construído manualmente não foi avaliado por métricas específicas, uma vez que a transição para modelos prontos foi realizada antes da análise detalhada do desempenho, entretanto os resultados iniciais indicaram a necessidade de explorar alternativas mais robustas, motivando a transição para modelos prontos disponíveis em bibliotecas como o scikit-learn, disponível em <https://scikit-learn.org/>

5.4 UTILIZAÇÃO DA BIBLIOTECA SCIKIT-LEARN PARA REGRESSÃO LOGÍSTICA

Após as fases de construção manual e implementação automatizada do modelo de Regressão Logística utilizando a biblioteca Scikit-learn, reconheceu-se a importância de lidar com desbalanceamento de classes para melhorar a capacidade preditiva. A escolha da técnica de oversampling, especificamente, foi adotada como uma estratégia eficaz para lidar com o desafio do desbalanceamento, proporcionando uma abordagem mais equilibrada na modelagem.

Durante a divisão dos dados em conjuntos de treinamento e teste, realizada pela função `train_test_split` da Scikit-learn, 20% dos dados foram reservados para teste, mantendo a configuração padrão dos hiperparâmetros do modelo nesse estágio.

A aplicação da técnica de oversampling visou mitigar os efeitos do desbalanceamento, que pode impactar negativamente o desempenho do modelo, especialmente em casos em que a classe minoritária é de interesse primordial.

Para avaliar o desempenho do modelo com a introdução das técnicas de balanceamento, foram utilizadas métricas padrão, incluindo acurácia, precisão, recall, F1-score, matriz de confusão e relatório de classificação, onde essas métricas forneceram uma visão abrangente da capacidade preditiva do modelo após a aplicação de técnicas de balanceamento.

A análise comparativa dos resultados entre o modelo original e o modelo com técnicas de oversampling destacou melhorias significativas, especialmente na precisão da classe

minoritária. Já que refletiram diretamente na capacidade do modelo de fazer previsões mais equilibradas e confiáveis.

Embora a implementação de técnicas de balanceamento tenha proporcionado benefícios notáveis, a generalização do modelo permanece um desafio, assim o uso de estratégias adicionais para evitar problemas de overfitting ou underfitting são cruciais, considerando a complexidade do conjunto de dados e a dinâmica das relações entre as variáveis.

Esta etapa enfatiza a importância não apenas de escolher modelos robustos, mas também de aplicar técnicas adequadas para lidar com o desbalanceamento de classes, otimizando assim a capacidade preditiva do modelo de Regressão Logística.

5.5 AJUSTE FINO DE HIPERPARÂMETROS NO MODELO DE REGRESSÃO LOGÍSTICA

O ajuste fino dos hiper parâmetros desempenha um papel crucial na otimização do desempenho do modelo de Regressão Logística, optou-se por utilizar a ferramenta “GridSearchCV” a qual pertence a biblioteca scikit-learn, para automatizar e realizar uma busca exaustiva nos diferentes conjuntos de hiper parâmetros, proporcionando uma ampla exploração do espaço de parâmetros.

Todos os hiper parâmetros foram submetidos ao ajuste fino, cada um desempenhando um papel significativo no comportamento e desempenho do modelo. Isso permitiu uma análise abrangente de como diferentes configurações afetam a eficácia do modelo.

A estratégia de busca foi definida por meio de uma grade de valores, abrangendo várias opções para cada hiper parâmetro, a abordagem permitiu examinar uma ampla variedade de combinações e identificar aquelas que melhor se ajustam aos dados em questão.

Os impactos na performance do modelo foram observados principalmente na busca por uma precisão otimizada, que ao ajustar os hiper parâmetros, o objetivo foi alcançar uma melhoria significativa na precisão sem sacrificar excessivamente outras métricas essenciais.

Diversas métricas de avaliação, como validação cruzada, acurácia, precisão, recall, F1-score, confusão e relatório de classificação, foram consideradas para analisar a eficácia do ajuste fino. proporcionando uma visão holística do desempenho do modelo em diferentes aspectos.

Os resultados após o ajuste fino foram comparados com os resultados anteriores do modelo construído manualmente e do modelo inicial com Scikit-learn. Observou-se uma

melhoria significativa na precisão, demonstrando a eficácia do ajuste fino em moldar o modelo de acordo com as características específicas dos dados.

As principais conclusões e insights obtidos destacam a importância dos hiper parâmetros na adaptação do modelo ao conjunto de dados, combatendo tanto o underfitting quanto o overfitting. O ajuste fino permitiu uma personalização refinada, maximizando a capacidade preditiva do modelo.

Os resultados do ajuste fino impactaram a escolha de modelos adicionais, estimulando a curiosidade em explorar novos algoritmos que pudessem elevar ainda mais as métricas de desempenho.

Em termos de contribuições para a tomada de decisões estratégicas, o ajuste fino dos hiper parâmetros proporcionou um modelo mais robusto e preciso, fornecendo insights valiosos para fundamentar escolhas relacionadas ao modelo de Regressão Logística.

Apesar dos desafios enfrentados, como encontrar hiper parâmetros que otimizassem a precisão sem prejudicar outras métricas, o ajuste fino foi uma etapa crucial para aprimorar o desempenho do modelo, onde a necessidade de ajustar o limiar de decisão de 0.5 para 0.65 destacou a importância de considerar a sensibilidade do modelo às diferentes decisões de classificação.

5.6 MODELO DE RANDOM FOREST

A etapa de construção e refinamento do modelo de Regressão Logística proporcionou insights valiosos, no entanto, para uma análise mais abrangente, decidimos explorar o Modelo de Random Forest. Esta escolha se deve à capacidade desse modelo de lidar eficientemente com conjuntos de dados complexos, oferecendo uma abordagem mais robusta e poderosa.

A Random Forest é uma técnica de aprendizado de máquina baseada em ensemble, que combina as previsões de vários modelos mais simples (árvores de decisão) para melhorar o desempenho geral.

Cada árvore de decisão é treinada de forma independente, tornando o modelo robusto a padrões específicos em determinadas subamostras. Introduzindo aleatoriedade durante a construção de cada árvore, como a escolha aleatória de características para divisões, ajuda a reduzir a correlação entre as árvores e aumenta a generalização do modelo.

Foram replicadas as etapas anteriores das divisões do conjunto de dados e do ajuste fino dos hiper parâmetros.

A avaliação do desempenho do modelo de Random Forest envolveu a aplicação de métricas padrão, incluindo acurácia, precisão, recall, F1-score, matriz de confusão e relatório de classificação. Essas métricas forneceram uma visão holística da capacidade preditiva do modelo em comparação com o modelo de Regressão Logística.

A análise comparativa entre os modelos revelou insights importantes sobre a natureza do conjunto de dados, o modelo de Random Forest demonstrou uma capacidade superior de lidar com complexidades e padrões não lineares, resultando em melhorias substanciais em métricas de desempenho, no entanto, é vital considerar os trade-offs, como a interoperabilidade reduzida em comparação com o modelo de Regressão Logística.

Apesar de notas excelentes, essas mesmas notas indicam om overfit, mesmo sendo tratadas pelo ajuste fino dos hiper parâmetros e com os resultados das métricas observadas, de uma maneira holística, as notas não denunciam exatamente um overfit. Uma forma de identificar que este modelo não está sofrendo com o sobre ajuste, é implementar um teste em um cenário real.

5.7 MODELO DE XGBOOST

Ao avançar em nossa análise, incorporamos o Modelo de XGBoost (Extreme Gradient Boosting), uma técnica sofisticada de gradient boosting, visando ainda mais aprimoramento da precisão do nosso modelo preditivo. O XGBoost é conhecido por sua eficácia em lidar com conjuntos de dados desafiadores e a capacidade de fornecer resultados superiores em uma variedade de tarefas de aprendizado de máquina.

O XGBoost oferece várias características que contribuem para sua eficácia e popularidade em competições de aprendizado de máquina, baseado no conceito de gradient boosting, o XGBoost treina uma série de modelos fracos (geralmente árvores de decisão) sequencialmente, melhorando iterativamente a precisão do modelo.

A implementação do Modelo de XGBoost foi realizada no ambiente Python, assim como nos modelos anteriores, dividimos os dados em conjuntos de treinamento e teste, reservando 20% dos dados para avaliação de desempenho.

O ajuste de hiper parâmetros foi uma etapa crucial para otimizar a precisão do modelo, explorou-se configurações diferentes, incluindo a taxa de aprendizado, o número máximo de profundidade das árvores e o peso das instâncias, a estratégia de busca utilizada foi o GridSearchCV, que nos permitiu testar várias combinações de hiper parâmetros para identificar a configuração mais eficaz.

Ao avaliar o desempenho do Modelo de XGBoost utilizando um conjunto abrangente de métricas, incluindo acurácia, precisão, recall, F1-score, matriz de confusão e relatório de classificação, essas métricas forneceram insights valiosos sobre a capacidade do modelo de fazer previsões precisas e identificar corretamente as classes de interesse.

Os resultados obtidos com o Modelo de XGBoost indicaram uma melhoria significativa nas métricas de desempenho em comparação com os modelos anteriores, assim a capacidade do XGBoost em lidar com relacionamentos não lineares e sua resistência ao overfitting destacaram-se, proporcionando uma solução mais robusta para nosso problema de classificação.

Considerando a complexidade do modelo, é importante ponderar os benefícios em relação à interoperabilidade, em cenários onde a interoperabilidade é crucial, pode ser necessário fazer escolhas estratégicas entre modelos mais complexos e modelos mais simples, como a Regressão Logística.

5.8 COMPARAÇÃO DE MODELOS

Ao longo deste processo analítico, construiu-se e explorou-se diferentes modelos preditivos, cada um com suas próprias características e complexidades. A Regressão Logística, implementada manualmente e por meio da biblioteca Scikit-learn, proporcionou uma visão inicial e compreensão das nuances dos dados. Em seguida, a incorporação do Modelo de Random Forest acrescentou robustez, superando desafios específicos, e finalmente, a introdução do Modelo de XGBoost demonstrou um salto significativo em termos de precisão e capacidade de generalização.

Regressão Logística: A simplicidade inicial da Regressão Logística revelou-se valiosa para uma visão preliminar dos dados, mas sua eficácia foi limitada diante de complexidades mais profundas.

Random Forest: A aplicação do Modelo de Random Forest introduziu um nível de complexidade superior, proporcionando uma interpretação mais profunda dos padrões nos dados e um aumento substancial na precisão.

XGBoost: O Modelo de XGBoost destacou-se como o mais eficaz em termos de precisão e generalização. Sua capacidade de lidar com relacionamentos não lineares e evitar overfitting contribuiu significativamente para um modelo mais robusto.

A avaliação de desempenho, com base em métricas como acurácia, precisão, recall e F1-score, foi fundamental para a tomada de decisões estratégicas, essas métricas permitiram

uma compreensão abrangente dos pontos fortes e fracos de cada modelo, auxiliando na escolha da abordagem mais adequada para nosso problema específico.

Além da comparação de modelos, foi crucial considerar a implementação prática desses modelos em ambientes de produção, a exportação e importação de modelos treinados são etapas essenciais nesse processo. Os modelos vencedores, como o Modelo de XGBoost, podem ser exportados utilizando bibliotecas como o joblib ou pickle, disponíveis em <https://joblib.readthedocs.io/>, <https://github.com/python/cpython/blob/3.12/Lib/>, e importados em sistemas de produção para realizar previsões em tempo real.

Eficiência: Modelos mais eficazes podem resultar em previsões mais precisas, otimizando processos e reduzindo custos operacionais.

Tomada de Decisões Holísticas: Com previsões mais precisas, as organizações podem tomar decisões mais seguras e estratégicas impactando positivamente seus resultados.

Adaptação Contínua: A implementação em produção permite a adaptação contínua dos modelos, incorporando novos dados e garantindo relevância ao longo do tempo.

Em resumo, a comparação de modelos não se limitou apenas à precisão, também levou em consideração a interpretabilidade, o custo computacional e a escalabilidade, onde cada modelo tem seu lugar em diferentes contextos, e a escolha foi guiada pelos objetivos específicos e requisitos do problema em questão. O Modelo de XGBoost destacou-se como uma escolha promissora para cenários em que a precisão é de suma importância, e a escalabilidade do modelo é viável no ambiente de produção.

6 RESULTADOS E DISCUSSÕES

No cenário financeiro, o risco de crédito se destaca como um desafio proeminente onde a inadimplência, representando a falha de um devedor em cumprir suas obrigações de pagamento, emerge como um dos subtipos mais evidentes desse risco, abordou-se o impacto multifacetado da inadimplência, desde perdas financeiras diretas para os credores até efeitos indiretos na reputação e rentabilidade das instituições financeiras.

Além da inadimplência, explorou-se outros desafios que impactam o mercado de crédito, incluindo taxas de juros elevadas, fraudes e regulamentações, pois taxas de juros elevadas aumentam os riscos de inadimplência e criam um ciclo de endividamento, enquanto fraudes geram perdas financeiras diretas para os credores e impactam negativamente os tomadores de crédito legítimos.

Em meio a esses desafios, a ciência de dados e a análise estatística surgem como soluções, onde modelos preditivos, análises avançadas e algoritmos de aprendizado de máquina oferecem insights valiosos para a identificação de tomador de crédito de alto risco e a previsão de inadimplências, assim a análise em tempo real permite decisões de crédito informadas, otimizando recursos e aprimorando a prevenção de fraudes.

A implementação de modelos preditivos é uma das contribuições mais marcantes, já que esses modelos utilizam algoritmos avançados para analisar vastos conjuntos de dados históricos e variáveis relevantes, proporcionando uma avaliação mais precisa do risco de inadimplência. Essa abordagem permite identificar padrões complexos de comportamento financeiro, auxiliando na previsão de potenciais problemas de pagamento.

A análise avançada de dados permite uma compreensão mais profunda das nuances do comportamento dos devedores, ao examinar uma ampla gama de variáveis, como histórico de crédito, capacidade de pagamento, e fatores socioeconômicos, as instituições financeiras podem obter insights mais refinados, onde não apenas melhora a precisão na identificação de devedores de alto risco, mas também contribui para estratégias mais personalizadas de concessão de crédito.

A aplicação de algoritmos de aprendizado de máquina oferece uma abordagem dinâmica na avaliação de riscos, esses algoritmos têm a capacidade de aprender e se adaptar com base em novos dados, refinando constantemente suas previsões, ao integrar essa capacidade de adaptação, as instituições financeiras podem melhorar a eficácia de seus modelos ao longo do tempo, acompanhando as mudanças nas tendências de comportamento financeiro.

A capacidade de realizar análises em tempo real é uma vantagem estratégica significativa, pois ao monitorar continuamente o comportamento de pagamento dos tomadores de crédito, as instituições podem identificar precocemente sinais de possíveis inadimplências, onde permite intervenções proativas, como o contato imediato com o devedor para discutir opções de resolução. Essa abordagem preventiva não apenas reduz o risco de inadimplência, mas também fortalece a relação entre a instituição financeira e o cliente.

A análise eficaz de dados também contribui para a otimização de recursos. Ao direcionar os esforços de concessão de crédito para os tomadores com menor risco, as instituições podem alocar recursos de maneira mais eficiente. Além disso, a análise estatística desempenha um papel vital na prevenção de fraudes, pois ao identificar padrões suspeitos e comportamentos inconsistentes através de análises rigorosas pode ajudar a mitigar perdas financeiras e preservar a integridade do mercado.

A facilidade que a biblioteca scikit-learn proporcionou no processo de desenvolvimento permitiu o treinamento de alguns modelos de predição, proporcionando encontrar um modelo que melhor se adequa ao conjunto de dados, permitindo por meio de algumas métricas decidir quais modelos podem ser o melhor para resolver o problema da análise de risco.

<p>Random Forest: Uma precisão quase perfeita, junto com outras métricas acima de 85% mostram um excelente desempenho do modelo. É importante ressaltar que nem sempre as métricas vão descrever exatamente a realidade, mesmo estando embasadas com técnicas estatísticas. Este modelo acabou se mostrando uma boa escolha em caso de produção.</p>	<pre> Random Forest Acurácia: 0.9334052252340571 Random Forest Precisão: 0.9740952792481365 Random Forest Recall: 0.8896211523283346 Random Forest F1-score: 0.9299437941525293 Random Forest Matriz de Confusão: [[50126 1199] [5594 45086]] Random Forest Relatório de Classificação: precision recall f1-score support 0 0.90 0.98 0.94 51325 1 0.97 0.89 0.93 50680 accuracy 0.93 102005 macro avg 0.94 0.93 0.93 102005 weighted avg 0.94 0.93 0.93 102005 </pre>
<p>Extreme Gradient Boost: Apesar da precisão está na casa dos 60% - 70%, este modelo obteve resultados sólidos, sendo mais generalista em relação as suas decisões. Suas métricas estão bem lineares, não demonstrando problemas no treinamento, podendo ser uma boa opção para verificar se realmente o modelo com maior precisão (random forest) terá o mesmo desempenho em um cenário real.</p>	<pre> XGBoost Acurácia: 0.6777510906328121 XGBoost Precisão: 0.6730070527890575 XGBoost Recall: 0.6834846093133385 XGBoost F1-score: 0.6782053666702561 XGBoost Matriz de Confusão: [[34495 16830] [16041 34639]] XGBoost Relatório de Classificação: precision recall f1-score support 0 0.68 0.67 0.68 51325 1 0.67 0.68 0.68 50680 accuracy 0.68 102005 macro avg 0.68 0.68 0.68 102005 weighted avg 0.68 0.68 0.68 102005 </pre>

<p>Regressão logística: Apesar de ter as menores notas dentre os modelos estudados, o modelo de Regressão logística também atingiu resultado sólidos, visando a generalização do modelo.</p> <p>Assim como o modelo de Gradient Boost, suas métricas estão bem lineares, e mesmo possuindo uma menor precisão, sua vantagem está na eficiência computacional, vendo que possui um algoritmo mais simples e menos complexo que os algoritmos dos demais modelos.</p>	<p>Regressão Logística Acurácia: 0.6046762413607176 Regressão Logística Precisão: 0.6029672056162122 Regressão Logística Recall: 0.5982438831886345 Regressão Logística F1-score: 0.6005962580351216</p> <p>Regressão Logística Matriz de Confusão: [[31361 19964] [20361 30319]]</p> <p>Regressão Logística Relatório de Classificação:</p> <table border="1"> <thead> <tr> <th></th> <th>precision</th> <th>recall</th> <th>f1-score</th> <th>support</th> </tr> </thead> <tbody> <tr> <td>0</td> <td>0.61</td> <td>0.61</td> <td>0.61</td> <td>51325</td> </tr> <tr> <td>1</td> <td>0.60</td> <td>0.60</td> <td>0.60</td> <td>50680</td> </tr> <tr> <td>accuracy</td> <td></td> <td></td> <td>0.60</td> <td>102005</td> </tr> <tr> <td>macro avg</td> <td>0.60</td> <td>0.60</td> <td>0.60</td> <td>102005</td> </tr> <tr> <td>weighted avg</td> <td>0.60</td> <td>0.60</td> <td>0.60</td> <td>102005</td> </tr> </tbody> </table>		precision	recall	f1-score	support	0	0.61	0.61	0.61	51325	1	0.60	0.60	0.60	50680	accuracy			0.60	102005	macro avg	0.60	0.60	0.60	102005	weighted avg	0.60	0.60	0.60	102005
	precision	recall	f1-score	support																											
0	0.61	0.61	0.61	51325																											
1	0.60	0.60	0.60	50680																											
accuracy			0.60	102005																											
macro avg	0.60	0.60	0.60	102005																											
weighted avg	0.60	0.60	0.60	102005																											

Quadro 2 – Análise do Desempenho dos Modelos Aplicados

Fonte: elaborado pelo autor.

Com as métricas avaliadas, pode-se observar o desempenho superior do modelo Randon Forest, em comparação aos demais, possuindo métricas bem equilibradas indicando uma boa generalização e notas superiores aos 85% evidenciando sua eficiência, podendo ser considerado uma das melhores opções.

A flexibilidade pela utilização da biblioteca Joblib, possibilitou a importação direta dos modelos em novos aplicativos e ambientes de análise. Essa abordagem modular simplifica a integração dos modelos em diferentes contextos, promovendo a aplicação eficiente das previsões de risco de crédito por meio de interfaces intuitivas, como sites e aplicativos móveis, que oferecem aos usuários uma experiência acessível e eficiente na análise de crédito em tempo real.

Este estudo destacou a convergência entre os desafios no mercado de crédito para pessoa física e as soluções proporcionadas pela ciência de dados e análise estatística, a implementação eficaz dessas abordagens não apenas mitiga riscos e aprimora a eficiência, mas também contribui para a estabilidade do mercado e a sustentabilidade das instituições financeiras. A evolução constante nesse campo é crucial para enfrentar desafios dinâmicos e fomentar um ambiente financeiro mais resiliente e equitativo, assim contribuição da ciência de dados e análise estatística, em constante evolução, representa um avanço significativo na gestão de riscos e na promoção de práticas mais sustentáveis e equitativas no setor financeiro, destacando a importância contínua da inovação e adaptação no cenário financeiro.

7 CONCLUSÃO

Este estudo revelou que o mercado de crédito para pessoa física enfrenta desafios complexos, incluindo inadimplência, taxas de juros elevadas, fraudes e regulamentações rigorosas, essas questões não apenas representam ameaças financeiras diretas para as instituições, mas também têm implicações significativas na rentabilidade, reputação e na capacidade de fornecer serviços de crédito de maneira eficaz.

Contudo, observou-se que a ciência de dados e a análise estatística emergiram como aliadas poderosas na superação desses desafios, a capacidade de criar modelos preditivos, aplicar análises avançadas e utilizar algoritmos de aprendizado de máquina proporcionam uma visão valiosa para avaliar riscos, prever inadimplências e tomar decisões de crédito mais informadas.

Além disso, a implementação eficaz dessas abordagens não se limita à mitigação de riscos, ela contribui para a eficiência operacional, promove práticas de concessão de crédito responsáveis e fortalece a conformidade regulatória, diante de um cenário financeiro dinâmico exige-se uma abordagem proativa, e a ciência de dados oferece um caminho promissor para lidar com estes desafios

Portanto, a contínua evolução e adoção da ciência de dados e análise estatística não apenas são fundamentais para a saúde a longo prazo das instituições financeiras, mas também desempenham um papel crucial na promoção de um mercado de crédito mais resiliente, transparente e acessível para todos os consumidores. Este estudo destaca não apenas os desafios, mas também o potencial transformador que a inovação nesse campo pode trazer para o setor financeiro, destacando a importância de se abraçar e integrar essas ferramentas avançadas para moldar o futuro do crédito para pessoa física.

8 AGRADECIMENTOS

Quero agradecer de coração à minha família e amigos, que estiveram ao meu lado durante toda essa jornada. O apoio de vocês foi essencial em cada desafio que enfrentei. Aos meus mentores, um agradecimento especial ao Professor Marildo Domingos da Silva. Suas orientações no mercado financeiro foram cruciais, sou grato por compartilhar seu conhecimento, oferecer dicas valiosas e me ajudar a entender as diversas aplicações da

tecnologia nesse contexto. Ao Professor José Alexandre Ducatti, meu sincero agradecimento por me introduzir ao mundo da ciência de dados. Sua orientação iluminou o caminho que escolhi seguir. Obrigado por transmitir não apenas conhecimento, mas também a paixão por essa área. Cada um de vocês desempenhou um papel importante em meu crescimento acadêmico e profissional. Obrigado por acreditarem em mim e por serem parte fundamental desta conquista.

9 REFERÊNCIAS

BLATT, Adriano. **Avaliação de Risco e Decisão de Crédito**: um enfoque prático. São Paulo: Nobel, 1999.

BEHNIA, Pouran et al. Random Forest classification for volcanogenic massive sulfide mineralization in the Rouyn-Noranda Area, Quebec. **Ore Geology Reviews**, p. 105612, 2023.

BRIGHAM, E. F.; GAPENSKI, L. C.; EHRHARDT, M. C. **Administração financeira**: teoria e prática. São Paulo: Atlas, 2001.

CAMARGO, Camila. **Planejamento financeiro pessoal e decisões financeiras organizacionais**: relações e implicações sobre o desempenho organizacional no varejo. 2007.

CAOQUETTE, John B. et. al. **Gestão do Risco de Crédito**: o próximo grande desafio financeiro. Rio de Janeiro: Ed. Qualitymark, 1999.

CHEROBIM; ESPEJO. **Finanças pessoais**: conhecer para enriquecer! São Paulo: Atlas, 2011.

FÁVERO, Luiz. **Manual de Análise de Dados** - Estatística e Modelagem Multivariada com Excel®, SPSS® e Stata®. 1. ed. atual. [S. l.]: GEN LTC, 2017.

FERRAZ, J. C. (2021). **A educação financeira e sua importância na gestão financeira pessoal**. Revista Ciência Da Sabedoria, 2(2). Disponível em: <https://revista.faciencia.com.br/index.php/rcs/article/view/53>. Acesso em: 12 nov. 2023.

GÉRON, Aurélien. **Mãos à obra: aprendizado de máquina com scikit-learn & tensorflow**. 2. ed. São Paulo: Alta Books, 2019.

GRUS, Joel. **Data Science do zero: Noções fundamentais com python**. 2. ed. atual. São Paulo: Alta Books, 2021. 416 p. v. 1.

HAUSSEN, Luiza Zimmer. **O mercado de crédito no Brasil durante a pandemia de COVID-19**. Trabalho de Conclusão de Curso (Graduação em Administração) - Escola de Administração, Departamento de Ciências Administrativas, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2022. Disponível em: <https://www.lume.ufrgs.br/handle/10183/256009>. Acesso em: 11 nov. 2023.

MATOS, M. T. de; CONDURÚ, M. T.; BENCHIMOL, A. C. **Interseções na produção científica da ciência da informação e ciência de dados**. Acervo, [S. l.], v. 35, n. 2, p. 1–18, 2022. Disponível em: <https://revista.an.gov.br/index.php/revistaacervo/article/view/1804>. Acesso em: 14 nov. 2023.

MCKINNEY, Wes McKinney. **Python para Análise de Dados: Tratamento de Dados com Pandas, NumPy & Jupyter**. 1. ed. atual. [S. l.]: O'Reilly Media, 2023.

MUELLER, Andreas; GUIDO, Sarah. **Introduction to Machine Learning with Python: A Guide for Data Scientists**. 2. ed. [S. l.]: O'Reilly Media, 2016.

PELEGRINA, Guilherme Dean; DUARTE, Leonardo Tomazeli; GRABISCH, Michel. **Uma nova proposta para a regressão logística com base na integral de Choquet**. In: ANAIS DO SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL, 2023, São José dos Campos. Disponível em: <https://proceedings.science/sbpo-2023/trabalhos/uma-nova-proposta-para-a-regressao-logistica-com-base-na-integral-de-choquet?lang=pt-br>. Acesso em: 18 dez. 2023.

SANTOS, J. O. **Análise de Crédito: empresas e pessoas físicas**. 2 ed. São Paulo: Atlas, 2003.

SECURATO, José Roberto. **Crédito - Análise e Avaliação do Risco**. 2ª ed. São Paulo: Sain Paul, 2012.

SECURATO, José Roberto. **Decisões Financeiras em Condições de Risco**. São Paulo: Atlas, 1996.

SIKANDER, Rahu; GHULAM, Ali; ALI, Farman. XGB-DrugPred: computational prediction of druggable proteins using eXtreme gradient boosting and optimized features set. **Scientific Reports**, v. 12, n. 1, p. 5505, 2022. Disponível em: <https://www.nature.com/articles/s41598-022-09484-3>. Acesso em: 11 nov. 2023.

SILVA, José Pereira da. **Gestão e Análise de Risco de Crédito**. 4ª ed. São Paulo: Atlas, 2003.

TRAPP, A. C. G. & CORRAR L. J., (2005). Avaliação e Gerenciamento do risco Operacional no Brasil: Análise de caso de uma instituição financeira de grande porte. **Revista Contabilidade Finanças**. – USP, São Paulo, nº 37, pág. 24 a 36, Jan./Abr. 2005. Disponível em: <https://www.scielo.br/j/rcf/a/FGgHThDLgdnCR7sCNGx6H4g/?lang=pt> Acesso em 05 nov. 2023