UNIVERSIDADE PRESBITERIANA MACKENZIE
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

Tharles Maicon Freire dos Santos

Desenvolvimento de um Modelo de Análise de Crédito Baseado em Dados de Redes Sociais

Projeto de Pesquisa apresentado ao Programa de Pós-Graduação em Computação Aplicada da Universidade Presbiteriana Mackenzie como parte dos requisitos para a obtenção do título em mestre em computação aplicada.

Prof. Dr. Leandro Augusto da Silva

São Paulo
2024

S231d Santos, Tharles Maicon Freire dos
Desenvolvimento de um modelo de análise de crédito baseado em dados de redes sociais / Tharles Maicon Freire dos Santos

1.6 MB

Dissertação (Mestrado profissional em computação aplicada) – Universidade Presbiteriana Mackenzie, São Paulo, 2024.

Bibliografia: f. xx - xx
Orientador: Prof. Dr. Leandro Augusto da Silva

1. Análise de crédito 2. Redes socials 3. Machine Learning 4. Big Data, 5. Privacidade 6. Inclusão financeira I. Silva, Leandro Augusto da, orientador II. Título

CDD 658.15

Bibliotecária responsável: Maria Gabriela Brandi Teixeira – CRB 8 / 6339

Tharles Maicon Freire dos Santos

DESENVOLVIMENTO DE UM MODELO DE ANÁLISE DE CRÉDITO BASEADO EM DADOS DE REDES SOCIAIS

Projeto de Conclusão de Mestrado apresentada ao Programa de Pós-Graduação em Computação Aplicada como parte das exigências para obtenção do título de Mestre em Computação Aplicada.

Aprovado em

BANCA EXAMINADORA

Prof. Dr. Leandro Augusto da Silva
Universidade Presbiteriana Mackenzie

Prof. Dr. Gustavo Scalabrini Sampaio
Universidade Presbiteriana Mackenzie

Prof. Dr. Alexandra Aparecida Souza
Instituto Federal de São Paulo

AGRADECIMENTOS

A conclusão deste trabalho representa uma etapa significativa na minha trajetória acadêmica e pessoal. Isso não seria possível sem o apoio de muitas pessoas, às quais gostaria de expressar minha profunda gratidão.

Agradeço a Deus pela força e sabedoria concedidas ao longo deste percurso e à minha família por sempre estar ao meu lado, apoiando e incentivando.

Ao meu orientador, Professor Doutor Leandro Augusto da Silva, pela sua orientação precisa e encorajamento. Suas valiosas sugestões foram fundamentais para a realização deste trabalho. Aos demais discentes do programa, pelos conhecimentos compartilhados e pelo apoio mútuo. Suas discussões enriqueceram meu desenvolvimento acadêmico.

Aos amigos, pelo apoio emocional e incentivo constante. Suas palavras de encorajamento foram fundamentais nos momentos desafiadores.

À Universidade Presbiteriana Mackenzie, pela estrutura e recursos disponibilizados, e a todos os funcionários que contribuíram para a conclusão deste trabalho.

Por fim, agradeço a todos que, direta ou indiretamente, colaboraram para a realização deste mestrado. Muito obrigado!

RESUMO

A análise de crédito tradicional enfrenta diversos desafios, incluindo a dependência de dados históricos, que podem não refletir adequadamente o risco futuro, e a dificuldade de capturar fatores qualitativos e comportamentais. Em complemento aos dados históricos, há oportunidades de uso de dados de redes sociais que podem indicar o comportamento financeiro dos indivíduos, como frequência de postagens, redes de amigos e interações sociais. Estes dados de redes sociais uma vez transformados em variáveis podem agregar maior precisão na análise de crédito. Neste sentido que esta pesquisa propõe desenvolver um modelo inovador de análise de crédito, utilizando dados de redes sociais. A pesquisa visa explorar a viabilidade da aplicação de dados de redes sociais na análise de crédito e as vantagens dessa abordagem em comparação aos métodos tradicionais, que se baseiam em dados financeiros históricos. Para tanto são construídos e validados modelos preditivos utilizando técnicas de aprendizado de máquina. Esses modelos são então avaliados quanto à sua capacidade de atender aos pilares da análise de crédito – Caráter, Capacidade, Capital, Colateral e Condições – da mesma forma que as análises tradicionais baseadas em dados históricos. Os resultados da pesquisa indicam que a inclusão de dados de redes sociais pode melhorar significativamente a precisão dos modelos de análise de crédito, oferecendo uma visão mais completa e atualizada sobre os solicitantes. Além disso, a pesquisa discute as implicações éticas e de privacidade no uso de dados de redes sociais para fins de análise de crédito, bem como as implicações regulatórias. Este estudo contribui para a área de análise de crédito ao introduzir uma nova perspectiva baseada em big data e aprendizado de máquina, promovendo uma abordagem mais justa e inclusiva para a concessão de crédito.

Palavras-chave: *Análise de crédito, redes sociais, aprendizado de máquina, big data, privacidade, inclusão financeira.*

ABSTRACT

Traditional credit analysis faces various challenges, including a reliance on historical data that may not adequately reflect future risk and the difficulty in capturing qualitative and behavioral factors. In addition to historical data, there are opportunities to use social network data that can indicate individuals' financial behavior, such as posting frequency, friend networks, and social interactions. Once transformed into variables, this social network data can add greater accuracy to credit analysis. This research proposes developing an innovative credit analysis model using social network data. The study aims to explore the feasibility of applying social network data in credit analysis and the advantages of this approach compared to traditional methods that rely on historical financial data. To achieve this, predictive models are built and validated using machine learning techniques. These models are then evaluated on their ability to meet the pillars of credit analysis – Character, Capacity, Capital, Collateral, and Conditions – in the same way as traditional analyses based on historical data. The research results indicate that the inclusion of social network data can significantly improve the accuracy of credit analysis models, providing a more comprehensive and up-to-date view of applicants. Additionally, the research discusses the ethical and privacy implications of using social network data for credit analysis purposes, as well as regulatory implications. This study contributes to the field of credit analysis by introducing a new perspective based on big data and machine learning, promoting a fairer and more inclusive approach to credit granting.

Palavras-chave: *Credit analysis, social network, Machine Learning, big data, privacy, financial inclusion.*

Lista de Tabelas

Lista de Figuras

Sumário

Capítulo 1

INTRODUÇÃO

Nos últimos anos o cenário econômico global tem passado por transformações significativas, impulsionadas principalmente pela rápida evolução tecnológica e pelo crescente uso de dados na tomada de decisões financeiras (KIM; SOHN, 2020). O avanço das tecnologias de informação e comunicação permite o acesso a grandes volumes de dados, conhecidos como Big Data, que podem ser utilizados para aprimorar a eficiência e a precisão de modelos de tomada de decisão em diversas áreas, incluindo a análise de crédito. No contexto do mercado de crédito, a análise precisa e eficiente da capacidade de pagamento dos indivíduos tornou-se ainda mais crucial, especialmente em um ambiente econômico caracterizado por rápidas mudanças e incertezas (KUMAR; SINHA, 2020).

Tradicionalmente, a análise de crédito baseia-se em informações financeiras históricas e no histórico de crédito dos solicitantes (KOCH; MACDONALD, 2000). Esse método tradicional utiliza dados como renda, patrimônio, histórico de pagamento e dívidas existentes para avaliar o risco de inadimplência. No entanto, essa abordagem enfrenta vários desafios, incluindo a dependência de dados históricos que podem não refletir adequadamente o risco futuro, a dificuldade em acessar informações precisas e atualizadas, e as limitações na captura de fatores qualitativos e comportamentais. Essas limitações podem resultar em decisões de crédito imprecisas e na exclusão de indivíduos que, apesar de terem potencial de pagamento, não possuem um histórico de crédito robusto (KOCH; MACDONALD, 2000).

Nesse cenário, a utilização de dados de redes sociais surge como uma abordagem

inovadora e promissora para a análise de crédito (KOCH; MACDONALD, 2000). As redes sociais, amplamente utilizadas por milhões de pessoas ao redor do mundo, oferecem uma vasta quantidade de informações sobre comportamentos, relações sociais e atividades cotidianas dos usuários. Essas informações podem complementar os métodos tradicionais de análise de crédito, proporcionando uma visão mais abrangente e atualizada sobre os solicitantes. Dados de redes sociais podem incluir padrões de interação social, frequência de postagens, redes de amigos, além de indicadores comportamentais e psicográficos que não são capturados pelos dados financeiros tradicionais (BOYD; ELLISON, 2007).

As empresas usam regras de negócios baseados em conjunto de critérios, políticas e procedimentos para avaliar a solvência e o risco de inadimplência dos solicitantes de crédito. Essas regras são baseadas em dados financeiros e comportamentais e ajudam a automatizar a decisão de concessão de crédito (HAND; HENLEY, 2001a).

O presente estudo propõe desenvolver um modelo de análise de crédito utilizando dados de redes sociais, visando explorar as vantagens e a viabilidade dessa abordagem. A pesquisa visa responder a questões fundamentais, como quais padrões de comportamento financeiro podem ser identificados e como eles se aplicam nos modelos tradicionais de crédito, a possibilidade de aplicação de modelos baseados exclusivamente em dados de redes sociais, as vantagens trazidas por essa análise e a viabilidade de substituição dos métodos tradicionais por uma abordagem inovadora. Além disso, o estudo analisará os impactos éticos e de privacidade envolvidos no uso de dados de redes sociais para fins de análise de crédito, bem como as implicações regulatórias e de conformidade com a legislação vigente.

A adoção de dados de redes sociais na análise de crédito pode representar um avanço significativo, proporcionando uma avaliação mais completa e dinâmica do risco de crédito. Estudos mostram que a incorporação de dados de redes sociais pode melhorar a precisão dos modelos de análise de crédito, oferecendo percepções sobre comportamentos e relações sociais que não são capturados pelos dados financeiros tradicionais (GIL, 2006). No entanto, é essencial implementar essa abordagem de maneira responsável, respeitando a privacidade dos indivíduos e as normas éticas e legais vigentes. A integração de dados de redes sociais pode democratizar o acesso ao crédito, permitindo que indivíduos sem histórico de crédito formal sejam avaliados de maneira justa e precisa, contribuindo assim

para a inclusão financeira e o desenvolvimento econômico sustentável (MDPI, 2024).

1.1 Problema de pesquisa

O problema central desta pesquisa é avaliar a eficácia e a viabilidade de modelos de análise de crédito baseados em dados de redes sociais, em comparação aos métodos tradicionais que utilizam dados financeiros históricos. A investigação visa determinar se os dados de redes sociais podem oferecer uma análise mais precisa e atualizada do risco de crédito.

A análise tradicional de crédito enfrenta desafios significativos, incluindo a dependência de dados históricos que podem não refletir adequadamente o risco futuro e a limitação na captura de fatores qualitativos e comportamentais. Esses métodos deixam frequentemente de considerar mudanças recentes na situação financeira e comportamental dos indivíduos, resultando em avaliações de risco desatualizadas ou imprecisas (KOCH; MACDONALD, 2000).

Os modelos baseados em dados de redes sociais prometem superar essas limitações ao fornecer informações adicionais sobre os comportamentos e interações sociais dos indivíduos. Esses dados podem incluir a frequência e o tipo de interações sociais, as redes de amizades e outros comportamentos que podem ser indicativos da responsabilidade financeira e da estabilidade emocional, aspectos cruciais na determinação do risco de crédito. Estudos demonstram que a incorporação de dados de redes sociais pode melhorar a precisão dos modelos de análise de crédito, oferecendo percepções que não são capturadas pelos dados financeiros tradicionais (WHARTON, 2024).

Um estudo da SpringerLink (2024) investiga se os dados extraídos de redes sociais podem ser utilizados para criar modelos preditivos mais precisos do que os métodos tradicionais. Tal pesquisa traz percepções importantes dos fatores para análise do problema, como a aplicabilidade de modelos em diferentes setores, categorias e classes sociais, por meio de análise de fatores como comportamento de consumo, interações sociais e frequência de atividades online.

Além disso, esta pesquisa avaliará se a utilização de dados de redes sociais pode resul-

tar em uma inclusão financeira mais justa, considerando, aspectos como acessibilidade ao crédito e equidade na avaliação de risco. A possibilidade de generalização desses modelos para diferentes contextos e populações também será analisada, com foco na robustez e adaptabilidade dos algoritmos utilizados.

Adicionalmente, é crucial considerar as implicações éticas e legais do uso de dados de redes sociais na análise de crédito. A privacidade dos dados dos usuários deve ser rigorosamente protegida, e a transparência na utilização desses dados é fundamental para garantir a confiança dos consumidores e a conformidade com as regulações vigentes (SCIENCEDIRECT, 2024). A viabilidade técnica e operacional desses modelos também será analisada, considerando as diferentes realidades de acesso às redes sociais e as limitações computacionais (WHARTON, 2024).

1.2 Pergunta de pesquisa

Esta pesquisa visa responder a três perguntas fundamentais relacionadas à aplicação de modelos de análise de crédito baseados em dados de redes sociais:

1. **Padrões e Comportamentos Financeiros**: Quais padrões e comportamentos financeiros podem ser identificados em dados de redes sociais e como eles podem ser integrados aos modelos tradicionais de análise de crédito?
2. **Vantagens da Análise de Dados de Redes Sociais**: Quais são as vantagens que a análise de dados de redes sociais traz para a análise de crédito?
3. **Viabilidade de Substituição da Análise Tradicional**: É viável substituir a análise de crédito baseada em dados financeiros tradicionais por uma baseada em dados de redes sociais?

Os métodos empregados nesta pesquisa traduzem dados qualitativos e quantitativos para abordar essas questões.

Para responder a primeira pergunta, utiliza-se os dados coletados das redes sociais para identificar variáveis que possam indicar comportamento financeiro. Para tanto utiliza-se Scorecards baseados em dados estatísticos de empresas acreditadas no mercado

financeiro, comparando-os com dados demográficos e comportamentais extraídos das redes sociais. Essa abordagem permitirá identificar as variáveis que melhor demostrem o comportamento financeiro do indivíduo.

A segunda questão é abordada por meio de uma análise quantitativa dos resultados, utilizando dados de redes sociais e métricas como acurácia e precisão. Essa análise permitirá avaliar as vantagens da utilização de dados de redes sociais na análise de crédito.

A terceira pergunta é investigada aplicando o modelo criado durante a pesquisa a diferentes nichos da sociedade. Essa abordagem permitirá avaliar a viabilidade de substituir a análise de crédito tradicional por uma baseada em dados de redes sociais.

Por fim, o desempenho dos modelos de aprendizado de máquina foi comparado para identificar o melhor subconjunto explicável, apresentando aos pesquisadores o algoritmo de classificação mais eficaz, bem como a melhor seleção de dados.

Esta pesquisa fornecerá percepções valiosas sobre a aplicabilidade e as vantagens da utilização de dados de redes sociais na análise de crédito, bem como a viabilidade de substituir a abordagem tradicional por uma abordagem baseada nesses dados.

1.3 Objetivos

O objetivo principal deste estudo é desenvolver e validar um modelo de análise de crédito utilizando dados de redes sociais. Acredita-se que dados provenientes de redes sociais possam oferecer uma visão complementar e, por vezes, mais atualizada sobre o comportamento e a capacidade de pagamento dos indivíduos, em comparação com os métodos tradicionais de análise de crédito. Para atingir esse objetivo, são estabelecidos os seguintes objetivos específicos:

1. **Identificar variáveis relevantes em dados de redes sociais para análise de crédito**.

    - Nesta etapa, serão analisadas dados de uma rede social para extrair variáveis que possam ser indicativas do comportamento financeiro dos usuários. Estas variáveis podem incluir, mas não se limitam a, frequência de postagens, engajamento com conteúdo financeiro, interações com instituições financeiras, e

sinais de estabilidade pessoal e profissional. Estudos anteriores, como o de Lee et al. (2020), demonstraram que variáveis de redes sociais podem ser preditores úteis em modelos de crédito.

2. **Construir e validar modelos preditivos utilizando essas variáveis**.

   - Com as variáveis identificadas, serão desenvolvidos modelos preditivos que utilizem técnicas de aprendizado de máquina para avaliar o risco de crédito. O processo de validação incluirá a utilização de conjuntos de dados de teste para garantir que os modelos desenvolvidos sejam robustos e capazes de generalizar para novos dados. A pesquisa de Franklin et al. (2020a) ressalta a importância de validar modelos preditivos com dados reais para assegurar sua eficácia.

3. **Comparar o desempenho do modelo proposto com os modelos tradicionais de análise de crédito** (BAIDEN, 2011b).

   - O desempenho do modelo baseado em dados de redes sociais será comparado com modelos tradicionais de análise de crédito, como aqueles que utilizam histórico de crédito e dados financeiros tradicionais. Esta comparação será realizada utilizando dados que assegurem os pilares de avaliação de crédito. O estudo de Baziden 2011b fornece uma referência útil para essa comparação, destacando a eficácia dos métodos tradicionais.

1.4 Justificativa

A justificativa para este estudo baseia-se na crescente importância dos dados de redes sociais na era digital e no potencial desses dados para fornecer percepções adicionais para a análise de crédito. A análise tradicional de crédito enfrenta desafios, como a dependência de dados históricos e a limitação em capturar fatores qualitativos e comportamentais (SINKEY, 2002a). A utilização de dados sociais pode superar essas limitações, proporcionando uma visão mais completa e atualizada sobre os solicitantes de crédito (CROUHY; GALAI; MARK, 2001a).

Além disso, a inovação no uso de big data e técnicas de Aprendizado de Máquina pode melhorar a precisão das análises e beneficiar tanto as instituições financeiras quanto os

consumidores (THOMAS; EDELMAN; CROOK, 2002a). A aplicação dessas tecnologias na análise de crédito pode resultar em modelos mais precisos e eficientes, reduzindo o risco de inadimplência e melhorando a acessibilidade ao crédito para mais indivíduos (ALENCAR, 2000).

1.5 Contribuição

Este estudo contribui significativamente para a área de análise de crédito, introduzindo uma nova perspectiva baseada em dados de redes sociais. Espera-se que os resultados forneçam percepções valiosas sobre a aplicabilidade e as vantagens dessa abordagem inovadora, além de oferecer uma alternativa viável e potencialmente mais eficaz aos métodos tradicionais de análise de crédito (MICHEL, 2005a).

A pesquisa também visa estimular futuras investigações sobre o uso de big data e aprendizado de máquina em diversas aplicações financeiras, promovendo avanços na forma como os dados são utilizados para tomar decisões mais informadas e precisas (SOARES et al., 2019).

Capítulo 2

REFERENCIAL TEÓRICO

2.1 Análise de crédito tradicional

A análise de crédito tradicional é um processo fundamental adotado por instituições financeiras para avaliar a capacidade de pagamento e a idoneidade de clientes, tanto pessoas físicas quanto jurídicas (SAUNDERS; CORNETT, 2007). Essa metodologia se fundamenta primordialmente em informações financeiras e no histórico de crédito do cliente (KOCH; MACDONALD, 2000). A análise de crédito tradicional baseia-se em uma série de etapas e critérios estabelecidos para avaliar o risco de inadimplência de um potencial tomador de empréstimo. Alguns dos principais aspectos considerados nessa abordagem incluem a análise de demonstrações financeiras e fluxo de caixa; a avaliação do histórico de crédito e pontualidade nos pagamentos; a verificação de garantias e colaterais oferecidos; a análise da capacidade de geração de renda e estabilidade financeira; e a avaliação do setor de atuação e condições macroeconômicas (CAOUETTE; ALTMAN; NARAYANAN, 1998). Essa metodologia tradicional tem sido amplamente utilizada por instituições financeiras ao longo de décadas, fornecendo uma base sólida para a tomada de decisões de crédito (SINKEY, 2002b).

Embora a análise de crédito tradicional seja uma abordagem consolidada, ela enfrenta alguns desafios e limitações importantes, como a dependência de informações históricas que podem não refletir adequadamente o risco futuro, a subjetividade na interpretação de dados e julgamentos humanos, a dificuldade em lidar com grandes volumes de dados

e complexidade crescente, e a incapacidade de capturar fatores qualitativos e intangíveis (CROUHY; GALAI; MARK, 2001b). Essas limitações impulsionam a busca por métodos complementares, como modelos de pontuação de crédito e técnicas de inteligência artificial, para aprimorar a precisão e eficiência da análise de crédito (THOMAS; EDELMAN; CROOK, 2002b).

Conforme o CFA Chartered Financial Analyst Institute (2023), o risco de crédito é definido como a possibilidade de perda decorrente do não pagamento integral e pontual dos juros e/ou do principal pelo mutuário. Por sua vez, a Serasa (2023a) descreve a análise de crédito como um procedimento crucial adotado pelas instituições financeiras para avaliar o perfil dos clientes, visando mitigar o risco de inadimplência e determinar a viabilidade de oferecer crédito e sob quais condições.

Nesse processo, as instituições financeiras recorrem frequentemente a fontes de informação como os birôs de crédito, incluindo empresas renomadas como Serasa, Boa Vista e SPC, para obter dados sobre restrições no nome e pontuação de crédito dos clientes. A pontuação de crédito, segundo a Boa Vista, é um indicador crucial do comportamento financeiro do cliente e influência diretamente na aprovação de crédito, sendo que uma pontuação mais alta aumenta a confiança da instituição na capacidade de pagamento do cliente (SPC Brasil, 2023a).

Uma fonte de informação é o Cadastro Positivo, mantido pelas instituições financeiras, que registra o histórico de pagamentos dos consumidores. Um histórico positivo nesse cadastro pode ampliar as chances de aprovação de crédito e resultar em condições de pagamento mais favoráveis para o cliente (SPC Brasil, 2023a).

De acordo com SPC Brasil (SPC Brasil, 2023a), a habilidade de calcular o risco de crédito é fundamental para empresas que buscam oferecer condições de pagamento flexíveis aos clientes. Esse cálculo vai além da prevenção de perdas financeiras imediatas, sendo crucial para garantir a saúde financeira a longo prazo da empresa. Ao compreender o risco associado a cada transação de crédito, a empresa pode estabelecer limites de crédito adequados, definir termos contratuais que descrevam claramente o risco assumido e evitar a concessão de crédito a clientes com alto potencial de inadimplência.

Em suma, a análise de crédito tradicional é um processo multifacetado que se vale de uma variedade de fontes de informação para avaliar o risco de inadimplência e determinar

a concessão de crédito de maneira responsável e segura (Serasa, 2023a).

2.1.1 5 C's da análise de crédito

No artigo "The 5 C's of Credit in the Lending Industry", Baiden (2011a) descreve os cinco principais fatores na concessão de crédito: Caráter, Capacidade, Capital, Condições e Garantias. O autor fornece uma descrição detalhada de cada um desses aspectos e também explora o processo de análise devida, essencial para o processo de empréstimo.

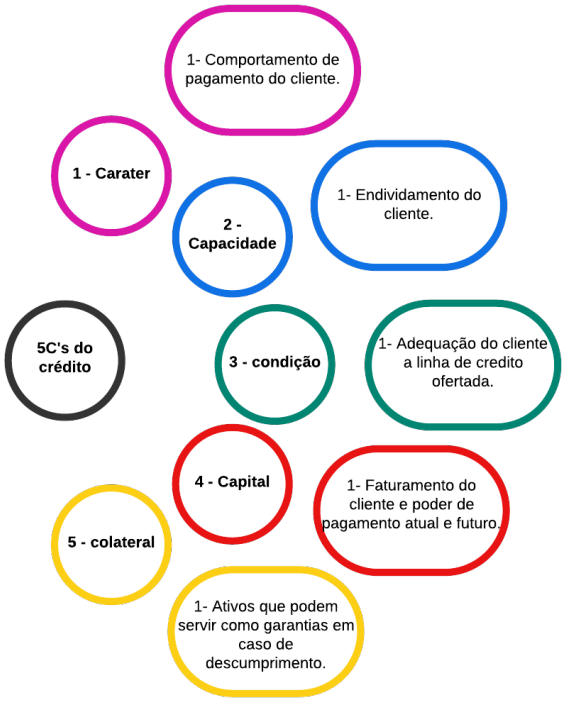

Figura 2.1: Os cinco C´s do crédito e a referência a análise de crédito.

Conforme a figura 2.1 cada pilar da análise de crédito representa fatores importantes na tomada de decisão, controle e recuperação dos dividendos.

Baiden (2011a) define **Caráter** como a disposição e determinação de cumprir uma obrigação de empréstimo. Ele exemplifica que, mesmo se o negócio do cliente enfrentar dificuldades financeiras, empresários de bom caráter se esforçarão para pagar um empréstimo. O caráter pode ser parcialmente confirmado analisando como o cliente pagou suas obrigações anteriores.

Capacidade, segundo Baiden (2011a), refere-se à habilidade da administração de

gerar fluxo de caixa suficiente para satisfazer todas as obrigações. Ele considera a capacidade como o aspecto mais crítico a ser avaliado. A análise da capacidade deve incluir quatro critérios principais: avaliar a fonte primária e a fonte alternativa de reembolso, considerar a qualidade e a fiabilidade do fluxo de caixa, e combinar revisões de informações financeiras de períodos anteriores com projeções futuras para demonstrar a saúde desse fluxo de caixa.

Baiden (2011a) descreve **Capital** como o patrimônio líquido que uma empresa possui para enfrentar problemas. O capital representa os fundos retidos na empresa, fornecendo uma proteção contra perdas inesperadas. Uma forte posição de capital é vista como garantia de que os proprietários da empresa permanecerão comprometidos com o negócio, mitigando riscos morais.

O quarto conceito, **Condições**, refere-se aos eventos econômicos nacionais, internacionais e locais, à indústria e ao próprio banco. As condições englobam o ambiente operacional do mutuário e também afetam o credor. Para o mutuário, as condições sob as quais opera podem influenciar significativamente a qualidade do crédito. A análise de crédito deve incluir uma avaliação da vulnerabilidade do mutuário às mudanças nas condições, como demanda, oferta, tecnologia, fornecedores, clientes, força de trabalho, normas trabalhistas e condições locais de emprego.

Finalmente, Baiden (2011a) aborda as **Garantias** como os ativos dados em garantia de um empréstimo. Empréstimos podem ser garantidos por contas a receber, estoques, equipamentos ou imóveis. Para reembolsar um empréstimo, a garantia deve ser liquidada, ou seja, transformada em dinheiro, superando, às vezes, obstáculos legais.

2.2 A inadimplência no Brasil

A inadimplência, caracterizada pelo não pagamento de dívidas no prazo acordado, é uma questão persistente no Brasil, tendo implicações econômicas e sociais profundas. Diversos fatores contribuem para a inadimplência no país. O desemprego é uma das principais causas, especialmente em momentos de crise econômica, como a recessão de 2015–2016 e a pandemia de COVID-19, que aumentaram a dificuldade dos consumidores em honrarem suas dívidas (IBGE, 2020).

Além disso, a falta de educação financeira é um fator crítico. A falta de conhecimento sobre gestão financeira leva a decisões de crédito inadequadas e, consequentemente, à inadimplência (Serasa, 2023b).

As condições de crédito oferecidas pelas instituições financeiras, como altas taxas de juros e prazos curtos, também contribuem para o problema. A facilidade de acesso ao crédito sem uma avaliação rigorosa da capacidade de pagamento do consumidor pode aumentar a inadimplência (SPC Brasil, 2023b).

A inadimplência tem consequências negativas tanto para os consumidores quanto para as instituições financeiras e a economia em geral. Para os consumidores, pode resultar em restrições de crédito, aumento das taxas de juros e problemas emocionais e de saúde devido ao estresse financeiro (CFA Institute, 2024). Para as instituições financeiras, representa perdas financeiras significativas, aumentando as provisões para devedores duvidosos e, em casos extremos, levando à insolvência (Banco Central do Brasil, 2023). Já para a economia, a inadimplência elevada reduz a disponibilidade de crédito, desacelerando o crescimento econômico e aumentando a volatilidade do mercado financeiro (The Economist, 2023).

Diversas estratégias podem ser adotadas para mitigar o problema da inadimplência no Brasil. Investir em programas de educação financeira pode auxiliar os consumidores a tomar decisões mais informadas sobre crédito e gestão financeira, reduzindo a probabilidade de inadimplência (Serasa, 2023b). Melhorar os processos de avaliação de crédito, utilizando modelos de pontuação de crédito mais precisos e abrangentes, também pode auxiliar as instituições financeiras a identificar melhor os riscos e conceder crédito de forma mais responsável (SPC Brasil, 2023b). Além disso, revisar as políticas de crédito, oferecendo condições mais favoráveis, como taxas de juros mais baixas e prazos de pagamento mais longos, pode facilitar o pagamento das dívidas pelos consumidores (Banco Central do Brasil, 2023).

Em suma, a inadimplência é um problema complexo e multifacetado no Brasil, que requer uma abordagem integrada para ser resolvido. Somente mediante uma combinação de educação financeira, melhores práticas de avaliação de crédito e políticas de crédito mais justas será possível reduzir a inadimplência e promover a saúde financeira dos consumidores e a estabilidade do sistema financeiro.

2.3 Serviços de proteção ao crédito.

Os serviços de proteção ao crédito são ferramentas essenciais para as instituições financeiras e os consumidores, por ajudarem a prevenir e mitigar o risco de inadimplência. Esses serviços incluem análise, classificação, monitoramento e proteção contra fraudes (Serasa, 2023b). A análise de crédito avalia a capacidade de pagamento do consumidor, enquanto a classificação de crédito classifica o risco de inadimplência. O monitoramento de crédito acompanha o histórico de pagamentos do consumidor, e a proteção contra fraudes detecta e previne fraudes de crédito (SPC Brasil, 2023b). Os serviços de proteção ao crédito oferecem vários benefícios, como redução do risco de inadimplência, melhoria da gestão de crédito, proteção dos interesses dos consumidores e redução dos custos de inadimplência (CFA Institute, 2024). No entanto, esses serviços também enfrentam desafios, como a complexidade da análise de crédito, limitações da avaliação do risco de inadimplência, dificuldade de detecção de fraudes e custos elevados para as instituições financeiras (Serasa, 2023b). Apesar dos desafios, os serviços de proteção ao crédito são fundamentais para as instituições financeiras e os consumidores, por ajudarem a prevenir e mitigar o risco de inadimplência. Eles oferecem benefícios significativos para a gestão de crédito e a proteção dos interesses dos consumidores (SPC Brasil, 2023b).

Os principais birôs de crédito no Brasil, segundo levantamento do Banco Neon (2024) são Serasa, SPC Brasil, Boa Vista Serviços e Quod.

1. **Serasa**: O Serasa é o birô de crédito mais conhecido do país e possui o banco de dados de pessoas e empresas mais completo. Ele oferece serviços como monitoramento completo de CPF com opção de bloqueio de consultas ao score, ofertas e feirões para negociação de dívidas, concessão de crédito, carteira digital e cadastro positivo (Serasa, 2023b).

2. **SPC Brasil**: O Serviço de Proteção ao Crédito (SPC Brasil) é um dos birôs mais antigos do país e faz parte da Confederação Nacional de Dirigentes Lojistas (CNDL). Em seu banco de dados tem informações enviadas por lojistas e comerciantes parceiros. Ele comercializa diversos pacotes de créditos para realizar consultas a CPFs e CNPJs em sua base, incluindo informações de renda presumida de pessoas, sugestões de limite de crédito, acesso ao quadro societário de empresas e dados detalhados so-

bre CPFs e CNPJs (SPC Brasil, 2023b).

3. **Boa Vista Serviços**: A Boa Vista Serviços é um birô de crédito relevante no Brasil, oferecendo serviços de análise de crédito e gestão de risco (VISTA, 2023). Com um crescimento acelerado principalmente após sua aquisição pela empresa norte-americana Equifax, Boa vista Serviços se consolidou como segundo principal birô de crédito do país acompanhando a Serasa.

4. **Quod**: O Quod é outro birô de crédito que opera no Brasil, fornecendo informações de crédito e serviços de análise de risco para instituições financeiras e empresas (QUOD, 2023).

Esses birôs de crédito desempenham um papel fundamental na economia brasileira, ajudando a proteger os negócios do risco de inadimplência e aumentando a segurança da economia em geral (Serasa, 2023b).

2.3.1 Negativação do devedor

A negativação do devedor se refere ao ato de cadastro do devedor nas listas de inadimplentes dos birôs de crédito. O Código de Defesa do Consumidor (CDC) estabelece diretrizes essenciais em relação à inscrição de clientes em listas de inadimplentes de serviços de proteção ao crédito. Embora não haja uma determinação específica sobre o prazo antes da inclusão de indivíduos na lista de inadimplentes, o CDC, em seu artigo 43, § 2º, estipula que o cliente deve ser formalmente notificado antes da inserção no cadastro de inadimplência. Essa responsabilidade recai sobre a entidade mantenedora do cadastro (Código de Defesa do Consumidor, 1990).

A jurisprudência do Superior Tribunal de Justiça (STJ) reforça essa obrigação através das súmulas 359 e 404, que afirmam respectivamente que cabe ao órgão mantenedor do Cadastro de Proteção ao Crédito notificar o devedor antes da inscrição e que é dispensável o aviso de recebimento (AR) na carta de comunicação ao consumidor sobre a negativação (Superior Tribunal de Justiça, 2023).

A ausência dessa comunicação prévia configura negativação indevida, passível de indenização por danos morais. O CDC também estipula que o credor tem até 5 dias úteis

para solicitar a exclusão do nome do consumidor do cadastro de inadimplentes. A falta de solicitação dentro desse prazo configura negativação indevida após os 5 dias úteis. Além disso, dividas com mais de 5 anos contados do vencimento da última parcela devem ser excluídas, conforme determina o Código de Defesa do Consumidor (1990).

Apesar de legalmente possível, a inclusão imediata do consumidor no cadastro de inadimplentes não é uma boa prática. Tal ação pode resultar em perda de confiança, deterioração da imagem da empresa e redução da fidelidade do cliente. Nesse contexto, a cobrança amigável é essencial. Um cronograma estruturado é fundamental para conduzir negociações amigáveis antes de tomar medidas mais severas, como a negativação do nome do devedor A negativação de clientes em serviços de proteção ao crédito deve seguir as diretrizes do CDC, respeitando prazos e procedimentos legais. A abordagem amigável na cobrança é recomendada, visando a recuperação de créditos de maneira eficiente e a preservação do relacionamento com os clientes. Essa prática não apenas facilita a recuperação de valores devidos, mas também mantém a reputação da empresa e a fidelidade do cliente (GROSSO, 2021). Finanças (2020) afirma que a abordagem amigável pode ser mais eficaz em recuperar dívidas rapidamente e manter um bom relacionamento com o cliente.

Um estudo realizado pela Boa Vista SCPC aponta que 67% dos clientes que passam por um processo de cobrança amigável têm mais chances de continuar a consumir produtos ou serviços da mesma empresa (SCPC, 2020). Evitar ações judiciais diminui significativamente os custos associados à cobrança, segundo a Serasa Experian, a cobrança judicial pode ser até cinco vezes mais cara do que a cobrança amigável, considerando taxas judiciais, honorários advocatícios e o tempo necessário para resolução do processo (EXPERIAN, 2019). A cobrança amigável envolve o uso de estratégias de comunicação assertivas e empáticas para incentivar o pagamento das dívidas sem recorrer a medidas punitivas ou à negativação do nome do devedor.

Na cobrança amigável, deve haver um cronograma estruturado para negociações amigáveis com acompanhamento constante e proativo para resolver a inadimplência antes da necessidade de medidas mais severas, como a negativação do nome do devedor. Esse processo inclui o envio de lembretes de pagamento, contato telefônico, propostas de renegociação e acompanhamento contínuo (SCPC, 2020; EXPERIAN, 2022). O tempo de negociação

amigável é crucial para a eficácia da cobrança extrajudicial. Grosso (2021) afirma que negociações realizadas em até 30 dias após o vencimento da dívida têm maiores chances de sucesso, com taxas de recuperação significativamente mais altas. Os primeiros dias são utilizados para implementar estratégias de cobrança amigável, durante os quais a comunicação proativa e as negociações têm altas chances de resolver a inadimplência \cite{scpc2020}. Após o período inicial de cobrança amigável, os próximos dias permitem uma reavaliação do risco e a implementação de estratégias mais intensivas, se necessário, garantindo que todas as tentativas amigáveis foram esgotadas antes de considerar o cliente como inadimplente crítico (LEE; CHOI, 2020; GROSSO, 2021). A pesquisa da Boa Vista demonstra que, após 90 dias de atraso, a probabilidade de recuperação da dívida diminui significativamente. Portanto, a marca dos 90 dias é um ponto crítico onde a previsão de inadimplência torna-se mais precisa e estatisticamente significativa (SCPC, 2022). A cobrança amigável pode ser mais eficaz em recuperar dívidas rapidamente. Estudos de Harvard (2018) indicam que estratégias de cobrança que utilizam abordagem personalizada e empática conseguem taxas de recuperação de até 80%, enquanto métodos punitivos ou impessoais ficam em torno de 50%.

2.4 Pontuação de crédito

A pontuação de crédito é uma métrica essencial no setor financeiro, amplamente utilizada para avaliar o risco de inadimplência de indivíduos e empresas. Este indicador é calculado com base em diversos fatores, incluindo histórico de pagamentos, nível de endividamento, tempo de crédito e outras informações financeiras relevantes (Serasa, 2023b). Desempenhando um papel crucial na tomada de decisões financeiras, a pontuação de crédito influencia diretamente processos como a concessão de empréstimos, aprovação de cartões de crédito, determinação de taxas de juros, e até mesmo na avaliação de risco para investimentos e seguros. Sua importância reside na capacidade de fornecer uma avaliação objetiva e quantitativa do risco associado a cada cliente (Serasa, 2023b).

A relação entre a pontuação e o risco de inadimplência é inversamente proporcional: quanto mais alta a pontuação, menor é o risco percebido de inadimplência, e vice-versa. Esta correlação torna a pontuação de crédito uma ferramenta valiosa em uma ampla

gama de situações financeiras, permitindo às instituições tomar decisões mais informadas e personalizadas em relação aos seus clientes e potenciais clientes (Serasa, 2023b). A pontuação de crédito é amplamente utilizada em diversas situações, incluindo:

1. **Aprovação de empréstimos e linhas de crédito**: a pontuação de crédito é um dos principais fatores considerados pelas instituições financeiras ao avaliar a concessão de empréstimos, financiamentos e linhas de crédito (EQUIFAX, 2021).

2. **Determinação das taxas de juros aplicadas**: clientes com pontuações de crédito mais altos geralmente têm acesso a taxas de juros mais baixas, enquanto aqueles com pontuações mais baixas enfrentam taxas mais elevadas, refletindo o risco associado (VISTA, 2023).

3. **Estabelecimento de limites de crédito**: a pontuação de crédito é utilizado para determinar os limites em cartões de crédito, empréstimos e outras linhas de crédito (EXPERIAN, 2022).

4. **Avaliação de risco para investimentos e seguros**: empresas de investimento e seguradoras utilizam a pontuação de crédito como uma das métricas para avaliar o risco de inadimplência de potenciais clientes (EXPERIAN, 2024).

No mercado financeiro, o cálculo da pontuação de crédito é realizado por birôs especializadas, como a Serasa Experian e a Boa Vista SCPC, que coletam e analisam informações de diversas fontes, como instituições financeiras, empresas de serviços públicos e órgãos governamentais (SCPC, 2022). Essas empresas utilizam algoritmos complexos e modelos estatísticos para atribuir um valor numérico a pontuação de crédito, variando geralmente de 0 a 1000 (EXPERIAN, 2022). Os principais fatores considerados no cálculo da pontuação de crédito incluem:

- **Histórico de pagamentos**: registros de atrasos, inadimplências ou quitações em dia são analisados, sendo que um histórico positivo de pagamentos pontuais contribui para uma pontuação mais alto (EQUIFAX, 2021).

- **Nível de endividamento**: A quantidade de dívidas em relação à renda é avaliada, pois um alto nível de endividamento pode indicar maior risco de inadimplência (EXPERIAN, 2024).

- **Tempo de crédito**: O tempo desde a abertura da primeira linha de crédito é considerado, já que um histórico mais longo de crédito pode indicar maior estabilidade financeira (SCPC, 2022).

- **Tipos de crédito**: A variedade de linhas de crédito, como empréstimos, cartões de crédito, financiamentos, entre outros, é analisada, pois uma diversidade de créditos pode indicar maior experiência e responsabilidade financeira (EXPERIAN, 2022).

- **Consultas recentes**: O número de consultas de crédito realizadas recentemente é considerado, pois muitas consultas em um curto período podem sugerir um risco mais elevado (EQUIFAX, 2021).

A pontuação de crédito tem um impacto significativo na vida financeira dos indivíduos e empresas. Uma pontuação alta pode abrir portas para melhores oportunidades de crédito, taxas de juros mais baixas e condições mais favoráveis em empréstimos e financiamentos (EXPERIAN, 2024). Por outro lado, uma pontuação baixa pode dificultar o acesso a crédito, resultar em taxas de juros mais altas e, em casos extremos, levar à negação de empréstimos ou linhas de crédito (SCPC, 2022). É importante monitorar regularmente a pontuação de crédito e adotar práticas financeiras saudáveis, como pagar contas em dia, manter um nível de endividamento equilibrado e evitar abrir muitas linhas de crédito em um curto período (EXPERIAN, 2024). Essas medidas podem ajudar a manter uma pontuação de crédito favorável e garantir melhores oportunidades financeiras no futuro.

2.5 Dados alternativos na análise de crédito

Com o crescente volume de dados disponíveis na internet, a análise de crédito se depara com uma vasta quantidade de informações sobre os requerentes, proporcionando uma base mais ampla e detalhada para a tomada de decisões. Esses dados provêm de diversas fontes e apresentam uma variedade de tipos, com diferentes formas de organização, seja estruturada ou não. Esse cenário demanda uma abordagem mais aprofundada e precisa na análise de crédito, ao passo que inviabiliza a análise manual em larga escala devido à sua magnitude. Nos últimos anos, a inteligência artificial (IA) emergiu como uma ferramenta

indispensável para as instituições financeiras que buscam se destacar em um mercado altamente competitivo. A aplicação da IA proporciona um aprimoramento significativo no desempenho e na precisão da análise de crédito. Por meio do uso de algoritmos de Aprendizado de Máquina e técnicas computacionais avançadas, a IA capacita as instituições a prever a inadimplência e calcular o risco de maneira mais precisa do que nunca. Dessa forma, a análise de risco não se limita mais ao histórico do requerente, mas é capaz de antecipar o comportamento futuro do devedor (ÓSKARSDóTTIR et al., 2019).

2.5.1 Big Data e técnicas de aprendizagem de máquina na análise de crédito

O estudo "O valor do big data para pontuação de crédito: melhorando a inclusão financeira usando dados de telefones celulares e análises de redes sociais" de (ÓSKARSDóTTIR et al., 2019) demonstra o poder do big data ao utilizar registros detalhados de chamadas. Conforme Óskarsdóttir et al. (2019), a combinação desses dados com informações tradicionais melhora significativamente o desempenho dos modelos de pontuação de crédito, especialmente em termos de lucro. Este exemplo de big data destaca como características de comportamento de chamadas podem prever a solvabilidade dos clientes de forma econômica e quase em tempo real, permitindo avaliações de crédito mais rápidas e precisas. Além disso, beneficia clientes com pouco ou nenhum histórico de crédito, melhorando a inclusão financeira e oferecendo percepções únicas sobre redes sociais e comportamento de reembolso.

Big Data pode melhorar a inclusão financeira e superar a pontuação de crédito tradicional, aproveitando fontes de dados não tradicionais e métodos de aprendizagem de máquina, mas a relevância dos dados e a prevenção de variáveis discriminatórias são cruciais (BAZARBASH, 2019).

Estes estudos sugerem que os grandes volumes de dados, a aprendizagem máquina e a Inteligência Artificial podem melhorar a precisão da pontuação de crédito, melhorando as avaliações dos clientes, o valor das garantias, as perspectivas de rendimento e reduzindo as perdas dos inadimplentes. Os dados alternativos devem cumprir com os 5v's do big data, Volume, Velocidade, Variedade, Veracidade e Valor, (ÓSKARSDóTTIR et al., 2019).

2.5.2 Análise de redes sociais

A utilização de dados de redes sociais tem sido estudada como uma ferramenta adicional para melhorar a precisão da análise de crédito. Vários estudos demonstram que essa abordagem pode ser altamente eficaz. Um estudo de Yao et al. (2024) conclui que a análise de dados de redes sociais pode prever o risco de inadimplência com uma precisão mais de 85%. Além disso, Lee e Kim (2020) afirma que a análise de dados de redes sociais pode identificar padrões de comportamento associados a um maior risco de inadimplência.

As informações pessoais e comportamentais extraídas de redes sociais podem ter um impacto significativo na análise de crédito. A presença de dados pessoais, como idade e gênero, pode ser um indicador valioso. Kumar e Gupta (2020) ressalta que essas informações podem prever o risco de inadimplência. Similarmente, Lee e Kim (2020) afirma que a frequência de postagens e a quantidade de amigos nas redes sociais são padrões comportamentais que podem ser correlacionados com um maior risco de inadimplência.

A análise de redes sociais é um campo em constante evolução, com aplicações em diversas áreas, incluindo a previsão de comportamentos financeiros. Uma das ferramentas mais importantes para a análise de redes sociais é a medida de centralidade, que ajuda a identificar os nós mais importantes na rede. A teoria de redes sociais baseia-se em conceitos como grau, intermediação e proximidade. O grau de um nó é o número de ligações que ele tem com outros nós; a intermediação refere-se à capacidade de um nó conectar outros nós; e a proximidade é a distância entre dois nós na rede (HAWE; WEBSTER; SHIELL, 2004; LAZEGA; HIGGINS, 2014).

A análise de redes sociais utiliza a álgebra de matrizes e a teoria dos grafos para mapear e interpretar a estrutura das redes. Isso inclui a criação de matrizes de adjacência e a distinção entre diferentes tipos de redes, como redes pessoais e redes completas, além de redes orientadas e não orientadas. Medidas de centralidade, como centralidade de grau, proximidade e intermediação, são essenciais para identificar os atores mais influentes em um sistema de rede. Essas medidas permitem entender melhor as dinâmicas sociais e os fluxos de informações e recursos nas redes (CAIANI; PARENTI, 2011; SILVA; RIBEIRO, 2016; EMIRBAYER, 1994).

Representações vetoriais densas de nós em uma rede, que capturam propriedades es-

truturais e relacionais dos nós para preservar as distâncias e proximidades na rede original. Essas representações permitem que técnicas de aprendizado de máquina tradicionais sejam aplicadas a dados de redes complexas. Representações vetoriais permitem modelar relações complexas entre nós, capturando tanto conexões diretas quanto indiretas. Isso é crucial para tarefas de predição de links, onde o objetivo é prever a existência de uma conexão futura entre dois nós com base na estrutura atual da rede (GROVER; LESKOVEC, 2016). Ao transformar nós em vetores em um espaço multidimensional, permite calcular a similaridade entre nós usando métricas como a distância euclidiana ou o cosseno da similaridade. Isso pode ser usado para recomendar novos amigos em redes sociais ou identificar potenciais parceiros de negócios em redes profissionais (PEROZZI; AL-RFOU; SKIENA, 2014). Com a representação vetorial, os nós podem ser classificados ou agrupados com base em suas características estruturais e relacionais. Na análise de crédito, clientes com perfis de conectividade semelhantes podem ser agrupados para avaliar riscos coletivamente, o que pode ser mais eficaz do que avaliar individualmente (TANG et al., 2015).

Outra técnica amplamente utilizada é a detecção de comunidades. A detecção de comunidades é o processo de identificar subgrupos em uma rede que são mais densamente conectados entre si do que com o resto da rede. Essa técnica revela a estrutura modular da rede, fundamental para entender dinâmicas sociais e comportamentos coletivos. Comunidades frequentemente representam grupos sociais ou funcionais em uma rede. Por exemplo, em redes sociais online, comunidades podem corresponder a grupos de amigos, colegas de trabalho ou pessoas com interesses comuns. Identificar essas comunidades pode ajudar a entender como informações e influências se propagam na rede (FORTUNATO, 2010). Comunidades bem definidas são úteis para analisar como a influência se espalha na rede. Por exemplo, influenciadores em uma comunidade podem ter um impacto desproporcional em relação à disseminação de informações ou comportamentos, o que é valioso para campanhas de marketing direcionadas (NEWMAN, 2010). Na análise de crédito, detectar comunidades de clientes interconectados pode ajudar a identificar clusters de risco sistêmico. Se um membro de uma comunidade entrar em inadimplência, há uma probabilidade maior de que outros membros do mesmo grupo também o façam, devido à interconexão e possíveis influências mútuas (BLONDEL et al., 2008).

2.5.3 Os pilares de crédito e dados de redes sociais

Com o advento das redes sociais, novos modelos de análise de crédito surgiram, aproveitando dados não convencionais para avaliar o risco. Ao utilizar dados de redes sociais na avaliação, busca-se cumprir os pilares da análise de crédito: capacidade, capital, condições, colateral e caráter.

A capacidade de um tomador de empréstimo de pagar suas dívidas é avaliada com base em sua renda, histórico de emprego e dívidas atuais. Informações de renda e declarações fiscais são os principais indicadores utilizados nos modelos tradicionais (SMITH, 2020). Modelos baseados em redes sociais podem inferir a capacidade de pagamento de um indivíduo por dados comportamentais e atividades online. Conforme Johnson (2019) informações sobre o emprego atual e padrões de gastos podem ser deduzidas de perfis e postagens nas redes sociais, complementando ou substituindo os dados financeiros tradicionais.

O capital refere-se aos ativos que um tomador possui, que podem ser utilizados para pagar a dívida em caso de inadimplência. Na análise tradicional são analisados os saldos bancários, investimentos, imóveis e outros ativos tangíveis (BROWN, 2018). Embora menos direta, a análise de redes sociais pode fornecer percepções sobre o capital de um indivíduo. White (2021) afirma que postagens sobre compras de alto valor ou propriedades podem indicar a posse de ativos significativos. Além disso, conexões e interações com instituições financeiras ou grupos de investimentos podem ser indicativos do capital disponível.

Para as condições econômicas gerais e a finalidade do empréstimo são avaliados fatores como a taxa de juros vigente, a economia do setor no qual o tomador está empregado e a finalidade específica do empréstimo (ADAMS, 2017). Dados de redes sociais podem ser utilizados para avaliar as condições econômicas e a percepção do tomador sobre o ambiente econômico. Discussões sobre eventos econômicos, mudanças de emprego e tendências setoriais nas redes sociais podem fornecer um contexto adicional sobre as condições em que o tomador está inserido (TAYLOR, 2019).

O colateral é qualquer ativo que o tomador oferece como garantia para o empréstimo. Este é um aspecto crítico em empréstimos de maior valor, onde imóveis ou veículos fre-

quentemente servem como colateral (MILLER, 2016). A análise baseada em redes sociais pode ser menos direta nesse aspecto, mas ainda pode oferecer informações valiosas. Postagens e fotos relacionadas a ativos, como casas ou carros, podem ser indicativos do colateral disponível. Além disso, a verificação de propriedade de bens valiosos por meio de registros públicos e menções nas redes sociais pode complementar a avaliação de colateral (DAVIS, 2020).

Tradicionalmente, o caráter do tomador é avaliado com base no histórico de crédito, comportamento passado em relação ao pagamento de dívidas e referências pessoais. Históricos de crédito e pontuações de crédito são os principais indicadores utilizados (WILSON, 2015). O caráter pode ser avaliado de forma mais holística através da análise de redes sociais. Atividades online, interações com outros usuários, consistência de comportamento e engajamento em comunidades virtuais podem fornecer uma visão abrangente do caráter de um indivíduo. Comentários positivos ou negativos, participação em atividades voluntárias e estabilidade de relacionamentos são indicadores valiosos que podem ser extraídos das redes sociais (CLARK, 2018).

A análise de crédito tradicional e a análise baseada em dados de redes sociais oferecem abordagens complementares. Enquanto a tradicional se baseia em dados financeiros concretos e históricos de crédito, a análise de redes sociais adiciona uma camada comportamental e contextual que pode enriquecer a avaliação do risco de crédito. Integrar ambas as abordagens podem proporcionar uma visão mais completa e precisa do perfil de risco dos tomadores de empréstimo, melhorando a eficácia e a inclusão financeira (CNUDDE et al., 2019).

2.6 Modelos preditivos de inadimplência

A inadimplência é um problema comum em muitas indústrias, incluindo a financeira. Os modelos preditivos de inadimplência são ferramentas importantes para prever a probabilidade de inadimplência de um cliente. Esses modelos podem ser baseados em técnicas de modelagem estatística e aprendizado de máquina e são amplamente utilizados para melhorar a gestão de risco de crédito (THOMAS; EDELMAN; CROOK, 2002b).

Com base na revisão da literatura existente, é possível descrever modelos de Ma-

chine Learning que são amplamente reconhecidos e utilizados em tarefas de classificação e predição, devido à sua eficácia comprovada e aplicabilidade prática em diversas áreas, incluindo a análise de crédito (HAND; HENLEY, 2001b; LESSMANN et al., 2015).

Regressão Linear e Regressão Logística

A regressão linear utiliza a relação entre variáveis para prever a probabilidade de inadimplência, assumindo uma relação linear e erros com distribuição normal (HOSMER; LEMESHOW; STURDIVANT, 2013). Já a regressão logística considera a probabilidade de inadimplência como uma variável binária. É amplamente utilizada em modelos de pontuação de crédito devido à sua capacidade de lidar com variáveis categóricas e binárias (COX, 1989).

Análise Discriminante

A análise discriminante classifica clientes em grupos de inadimplentes e não inadimplentes, assumindo distribuição normal multivariada das variáveis e matrizes de covariância iguais para todos os grupos (FISHER, 1936).

Máquinas de Vetores de Suporte (SVM)

SVMs são modelos discriminativos que buscam encontrar um hiperplano que melhor separa as classes de dados. São eficazes em espaços de alta dimensão e visam encontrar o hiperplano que melhor separa clientes inadimplentes dos não inadimplentes, considerando a inadimplência como uma variável binária (CORTES; VAPNIK, 1995).

Árvores de Decisão e Random Forest

Árvores de decisão são modelos baseados em regras que particionam o espaço de características em regiões homogêneas (QUINLAN, 1986). Random Forests combinam várias árvores de decisão para melhorar a robustez e reduzir os outliers (BREIMAN, 2001). Utilizam a técnica de ensemble learning que combina várias árvores de decisão para melhorar a precisão da previsão de inadimplência.

Redes Neurais

Redes neurais, especialmente aquelas com múltiplas camadas (deep learning), têm mostrado grande potencial em capturar padrões complexos nos dados. Modelam relações não lineares entre variáveis para prever a inadimplência, considerando-a como uma variável

binária. São amplamente utilizadas em modelos de pontuação de crédito (GOODFELLOW; BENGIO; COURVILLE, 2016).

Capítulo 3

METODOLOGIA

Neste capítulo, descreve-se o design da pesquisa e o framework utilizado, detalhando as metodologias aplicadas e as etapas seguidas no desenvolvimento da pesquisa. Apresenta-se, também, a definição e o escopo do modelo de análise de crédito, destacando os principais componentes que deram formato ao modelo pretendido, os dados e métodos utilizados no seu desenvolvimento, e as expectativas em relação ao modelo, definindo seu escopo, como também a forma de validação do modelo. Em seguida, apresenta-se a justificativa sobre a origem dos dados, destacando sua relevância e heterogeneidade. Com isso, detalha-se o passo a passo de coleta, processamento e análise dos dados, abordando as ferramentas utilizadas e os procedimentos adotados. Além disso, explicam-se as abordagens metodológicas e a forma de validação dos resultados pretendidos.

3.1 Design de pesquisa e framework

Os métodos de pesquisa têm importante papel em um trabalho científico, e o uso de diferentes métodos de análise nas ciências ajuda a construir um trabalho mais completo e de qualidade. Quando se fala em pesquisa, é preciso entender que é possível trabalhar para a construção de conhecimentos em muitas áreas, e que em cada uma delas pode ser necessária a adoção de diferentes critérios para que se alcancem os objetivos, que também podem ser variados (SILVA; NóBREGA, 2018).

Este trabalho visou avaliar majoritariamente os resultados e as perguntas de pesquisa

quantitativamente, utilizando métricas e comparativos numéricos. A pesquisa aplicou os conhecimentos levantados na criação de um modelo de crédito a ser utilizado pela sociedade, buscando construir um procedimento capaz de validar os conceitos abordados e sua generalização, criando regras e definições de fácil reprodução em diferentes indivíduos. Enquanto isso, qualitativamente, a pesquisa validou o atendimento do modelo aos pilares de crédito através da análise das variáveis obtidas.

A análise das variáveis em relação aos pilares do crédito (Caráter, Capacidade, Capital, Colateral e Condições) foi predominantemente qualitativa. Baseou-se em uma revisão de literatura e nos comentários de outros pesquisadores, identificando a relevância e a contribuição de cada variável aplicada no modelo para os pilares do crédito. Como a pesquisa não utilizou dados transacionais históricos para a aplicação de modelos tradicionais, essa abordagem mista proporcionou uma avaliação abrangente do modelo de análise de crédito, combinando percepções teóricas e práticas com dados numéricos para obter uma visão completa e bem fundamentada.

Segundo Michel (2005b), a pesquisa quantitativa é um método de pesquisa social que utiliza a quantificação nas modalidades de coleta de informações e no seu tratamento, mediante técnicas estatísticas, tais como percentual, média, desvio padrão, coeficiente de correlação, análise de regressão, entre outros.

De acordo com Alencar (2000), na perspectiva positivista, a investigação científica destaca a importância do teste de validade de uma hipótese pela experimentação, cujo objetivo maior é medir ou quantificar a extensão pela qual uma relação causa-efeito existe. Os cientistas dessa concepção teórica acreditam que os métodos utilizados pelas ciências naturais podem ser aplicados aos estudos da vida social, a qual seria, portanto, mensurável e quantificável, tendo o pesquisador à sua disposição dados estatísticos (evidências empíricas) para explicar a realidade social.

A estratégia geral de análise da pontuação de crédito consistiu em executar algoritmos de classificação e de regressão para a análise da probabilidade do indivíduo pertencer à classe inadimplente. O passo inicial foi a coleta dos dados das redes sociais definidas, organizando esses dados, definindo-os em estrutura relacional e de grafos. Logo após, foram executados modelos e algoritmos já existentes de análise de postagens e de coleta de dados de diligência comparados com dados estatísticos do mercado financeiro. Após

definir o risco de inadimplência atribuído ao indivíduo, avaliou-se a régua de corte para aprovação ou reprovação do crédito.

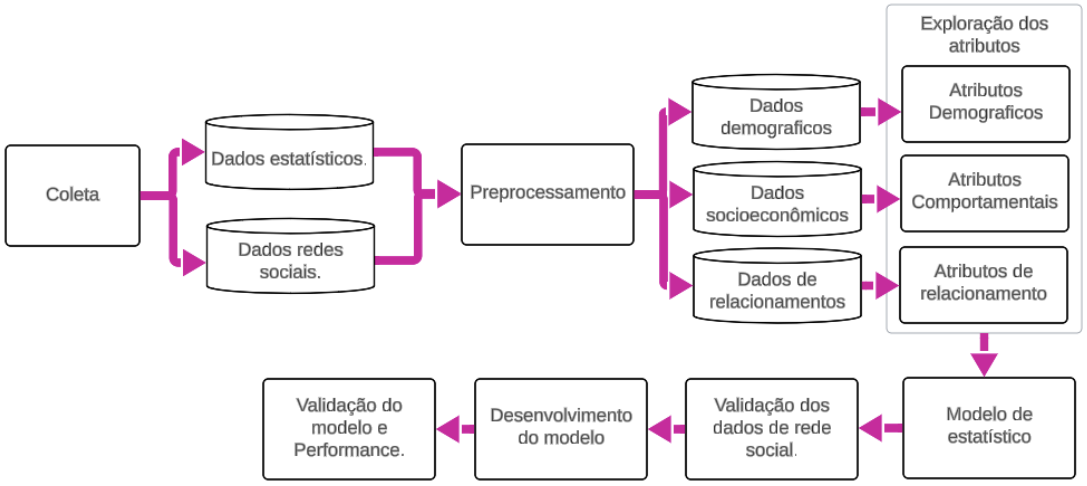

Figura 3.1: Design de pesquisa.

Conforme apresentado na Figura 3.1, a primeira etapa deste estudo envolveu a coleta de dados de duas fontes principais: dados estatísticos e dados de redes sociais. Os dados estatísticos foram extraídos da pesquisa Boa Vista de Inadimplência de 2022, que fornece um panorama detalhado sobre o comportamento de inadimplência no Brasil.

Os dados de redes sociais foram coletados a partir dos perfis de Facebook. Esses dados incluíram informações sobre relacionamentos, interações sociais, frequência de postagens, engajamento com conteúdos e outros indicadores de comportamento online.

No pré-processamento dos dados, as informações coletadas foram organizadas e padronizadas para garantir a consistência e a comparabilidade. Este processo envolveu:

- **Estatísticas da pesquisa Boa Vista:** foram identificados os dados da pesquisa mais relevantes para o entendimento do perfil do inadimplente.
- **Dados Demográficos:** Normalização de informações como idade, sexo, estado civil e localização.
- **Dados Socioeconômicos:** Padronização de variáveis relacionadas a emprego e educação.

- **Relacionamentos:** Extração e formatação de dados de redes sociais, incluindo número de amigos, indivíduos seguidos e seus seguidores.

Uma vez pré-processados, os dados foram submetidos a uma análise para identificar padrões e tendências relevantes. Esta fase envolveu a utilização de técnicas estatísticas para descrever e visualizar a distribuição das variáveis, bem como para identificar possíveis correlações entre os diferentes atributos.

Com base nos dados estatísticos obtidos, seleciona-se as variáveis mais relevantes para o processo de análise e se cria um Scorecard. O Scorecard é aplicado nos dados extraídos das redes sociais em busca de avaliar a propensão à inadimplência do indivíduo e por conseguinte se faz a análise discriminante para a identificação da classe inadimplente. Foi feita a análise de todas as variáveis extraídas da rede social, selecionando-se as variáveis a serem incluídas no modelo com base na sua relevância estatística e no seu poder preditivo.

Para garantir a qualidade e a validade dos dados de redes sociais, realizam-se testes de validação. Esses testes incluíram a verificação da consistência dos dados, a análise de overfit (sobreajuste) e a confirmação de que as informações coletadas eram representativas e relevantes para o estudo.

A etapa seguinte envolveu o desenvolvimento do modelo de análise de crédito. Para tanto se usam técnicas de Machine Learning, como regressão, classificação, árvores de decisão e redes neurais, para construir um modelo preditivo capaz de identificar potenciais inadimplentes com base nos dados coletados. O desenvolvimento do modelo incluiu a divisão dos dados em conjuntos de treinamento e teste, a seleção de hiper parâmetros e a otimização do modelo para maximizar a sua acurácia.

Finalmente, o modelo desenvolvido foi validado utilizando o conjunto de dados de teste. O desempenho do modelo foi avaliado com base em métricas como acurácia, precisão, recall e área sob a curva ROC. Além disso, realiza-se análise de sensibilidade para identificar a robustez do modelo e garantir que ele fosse aplicável em diferentes contextos e populações.

Faz-se também análises da capacidade do modelo em prever a pontuação de crédito baseada na propensão do indivíduo à inadimplência e da capacidade do modelo de prever corretamente futuros devedores inadimplentes quando complementado com a variável de

pontuação social. Além disso, se determina em quais grupos de clientes o uso de recursos sociais agrega mais valor. Por fim, o efeito econômico associado ao uso do score de crédito na previsão de inadimplência de empréstimos é discutido.

Existem vários desafios que podem impedir a concretização da estratégia descrita acima, como a construção da base relacionada à disponibilidade e limitações das APIs de redes sociais usadas, o uso de dados pessoais e as implicações legais, consentimentos e riscos de segregação e discriminação de indivíduos.

3.2 Definição e escopo do modelo

O objetivo do modelo desenvolvido é avaliar o risco de inadimplência dos solicitantes de crédito utilizando dados obtidos de redes sociais. Esse modelo identificará parâmetros que representem a propensão do indivíduo a deixar de pagar suas dívidas, baseando-se diretamente em estatísticas de mercado que representem o perfil do inadimplente. Desta forma, o modelo não utiliza nenhum dado transacional do mercado.

O modelo define como fonte de dados o perfil do indivíduo no Facebook, aliado às percepções da pesquisa Boa Vista de inadimplência sobre o perfil do endividado. A partir dos dados obtidos do perfil do Facebook, serão extraídos elementos para identificar quais os parâmetros relevantes para a análise do comportamento financeiro e compará-los ao perfil dos indivíduos inadimplentes, conforme a pesquisa Boa Vista.

O grande desafio inicial é aplicar os dados estatísticos sobre o conjunto de dados do Facebook. Técnicas estatísticas serão utilizadas para integrar os dados percentuais aos dados do Facebook. O uso de proporções, WoE (Weight of Evidence) e IV (Information Value) será empregado para representar um valor de pontuação dos dados estatísticos e aplicá-los ao conjunto de dados.

Os dados estatísticos contribuem proporcionalmente para a probabilidade de um indivíduo deixar de cumprir com seus pagamentos conforme estipulado. Assim, técnicas computacionais e de análise discriminante serão aplicadas para definir uma régua e corte que permita a classificação binária dos indivíduos como adimplentes ou inadimplentes.

Durante o desenvolvimento do modelo, práticas de coleta e armazenamento dos dados,

em combinação com técnicas de engenharia de dados, foram aplicadas como parte do processo de coleta e análise dos dados. Fontes adicionais de dados poderão ser utilizadas para enriquecer as informações extraídas dos perfis.

Na modelagem ainda se aplica técnicas de balanceamento do conjunto de dados para garantir que ele represente objetivamente ambas as categorias da variável alvo. Os dados serão divididos em conjuntos de treinamento e teste, e técnicas de validação cruzada serão aplicadas para evitar overfitting. Os hiper parâmetros serão ajustados para otimizar o desempenho dos modelos computacionais.

Métricas de validação, como acurácia, matriz de confusão, coeficiente de Gini, teste de Kolmogorov-Smirnov (KS) e curva ROC, serão definidas para avaliar o desempenho dos modelos. As probabilidades preditivas serão convertidas em scores de crédito para indicar o nível de risco. Thresholds específicos serão definidos para categorizar os solicitantes em grupos de risco e estabelecer uma régua de corte para a definição de adimplentes e inadimplentes.

Para a concessão do crédito, será definido um threshold que determinará o ponto de corte ajustado ao objetivo do negócio.

3.3 Scorecard boa Vista

O Scorecard de crédito representa de forma mais compreensível a pontuação de crédito atribuída ao indivíduo. No âmbito financeiro, é uma boa prática representar a propensão a inadimplência em forma de Scorecards. Nesses Scorecards, o valor da evidência é transformado em pontos em uma escala.

Para o desenvolvimento, foram utilizados dados da Pesquisa de Inadimplência e Endividamento realizada pela empresa Boa Vista (2022), subsidiária brasileira da Equifax, um dos maiores bureaus de crédito dos Estados Unidos, realizada durante o primeiro semestre de 2022. A Boa Vista é reconhecida como um dos principais birôs de crédito do Brasil e conduz regularmente estudos sobre inadimplência e endividamento para compreender melhor as dinâmicas financeiras da população. A pesquisa aborda diversos aspectos relacionados à inadimplência e ao endividamento no Brasil. Foram coletados dados demográficos

dos perfis dos devedores, investigou as razões para inclusão nas listas de restrições dos bureaus de crédito, examinou a relação de endividamento na sociedade e seu impacto, identificou as principais razões para a inadimplência e explorou estratégias de reestruturação de dívidas e alternativas para evitar a inadimplência. Esses dados são valiosos para entender as características e os comportamentos dos devedores brasileiros, além de fornecer percepções sobre os desafios enfrentados por eles. A pesquisa da Boa Vista é uma fonte confiável de informações sobre inadimplência e endividamento, contribuindo para uma análise abrangente da situação financeira no país (Boa Vista, 2022).

A pesquisa distingue a inadimplência e o endividamento. "inadimplência" é definido como o ato de não pagar dívidas em um prazo determinado, enquanto "endividado" refere-se a dívidas pendentes que ainda não estão vencidas.

A pesquisa foi conduzida de forma online com uma amostra de 1500 participantes de todas as regiões do Brasil. As características demográficas e socioeconômicas dos participantes são apresentadas abaixo:

- **Distribuição de Gênero:**
  - Participantes masculinos: 45%
  - Participantes femininos: 55%
- **Situação Financeira:**
  - 28% dos participantes relataram estar desempregados.
- **Distribuição Geográfica:**
  - 70% dos participantes residiam nas regiões sul e sudeste do Brasil.
  - 30% dos participantes residiam nas regiões norte, nordeste e centro-oeste.
- **Distribuição de Idade:**
  - 27% dos participantes tinham entre 18 e 30 anos.
  - 35% dos participantes tinham entre 31 e 40 anos.
  - 23% dos participantes tinham entre 41 e 50 anos.
  - 10% dos participantes tinham entre 51 e 60 anos.

- 5% dos participantes tinham mais de 60 anos.

- **Estado Civil:**
  - 54% dos participantes eram casados.
  - 35% dos participantes eram solteiros.
  - 10% dos participantes eram divorciados.
  - 1% dos participantes eram viúvos.

- **Classe Socioeconômica:**
  - 3% dos participantes pertenciam à classe socioeconômica alta (A-B).
  - 25% dos participantes pertenciam à classe socioeconômica média (C).
  - 72% dos participantes pertenciam à classe socioeconômica baixa (D-E).

A pesquisa também foca nos motivos principais para negativação e o tempo de atraso das dívidas. Conforme a pesquisa, a maioria dos entrevistados tem dívidas com atrasos superiores a 90 dias e em muitos casos mais de um débito pendente.

- **Tempo de atraso das contas:**
  - 86% mais de 90 dias.
  - 10% entre 30 e 90 dias.
  - 4% menos de 30 dias

Ao extrair os dados percentuais é possível utilizar estes dados para relacionar o perfil do indivíduo a propensão de deixar de pagar suas dívidas por um período. Os principais motivos para a negativação e o tempo de atraso das dívidas foram analisados. Conforme a pesquisa, a maioria dos entrevistados tem dívidas com atrasos superiores a 90 dias e em muitos casos mais de um débito pendente.

A pesquisa traz dados para analisar o perfil dos devedores e identificar características de risco de possíveis inadimplências. Ao analisar os dados demográficos dos participantes, o Scorecard identifica indivíduos com maior probabilidade de restrição de crédito, resultando na inclusão na lista de restrição de crédito.

3.3.1 Dados e Pré-Processamento

Variáveis Utilizadas O conjunto de dados consiste em várias variáveis que representam diversas características demográficas e socioeconômicas dos participantes. As variáveis e suas descrições são as seguintes:

Variável	Descrição
gender	O gênero dos participantes
age	A faixa etária dos participantes
employed	Indica se o participante está financeiramente ativo
region	A localização demográfica do participante
marital_status	Estado civil do indivíduo

Tabela 3.1: Descrição das variáveis - Scorecard Boa Vista

A tabela 3.1 apresenta o nome e a descrição de cada uma das variáveis analisadas.

Definição da variável dependente e variáveis independentes

A **variável dependente**, também definida como variável alvo, é a variável que se prevê. Neste conjunto de dados, define-se como variável alvo o tempo de inadimplência superior a 3 meses, levando à restrição de crédito. A escolha dos 90 dias de atraso como ponto de ruptura baseia-se em dados empíricos e práticas de mercado, particularmente na pesquisa realizada pela Boa Vista 2022. Esta pesquisa indica que após 90 dias de atraso no pagamento, a probabilidade de recuperação da dívida diminui significativamente, tornando-se um indicador crucial de inadimplência. Ao incorporar a prática de cobrança amigável na justificativa para a escolha dos 90 dias, alinhamos a abordagem do Scorecard com as práticas operacionais e de mercado, fortalecendo a validade da escolha e garantindo que o Scorecard seja não apenas teoricamente sólido, mas também pragmaticamente relevante.

3.3.2 Análise e processamento

As variáveis foram analisadas com base no seu poder de diferenciar entre eventos binários em uma variável alvo (Variável dependente.). O Índice de Informação (**Information**

Value - IV) é uma métrica usada para medir a força de uma variável preditora na diferenciação entre eventos binários (inadimplência maior que 90 dias/ menor que 90 dias). A interpretação do IV é realizada conforme os seguintes critérios:

- - **IV $<$ 0,02**: Previsor Não Útil
- - **0,02 $\leq$ IV $<$ 0,1**: Previsor Fraco
- - **0,1 $\leq$ IV $<$ 0,3**: Previsor Médio
- - **IV $\geq$ 0,5**: Previsor Muito Forte

Feature Name	**IV**	**IV Evaluation**
age	0,64	preditor muito forte
employed	0,41	preditor forte
gender	0,02	preditor fraco
marital_status	0,42	preditor forte
region	0,33	preditor forte

Tabela 3.2: Avaliação do IV das variáveis do Scorecard Boa Vista

A tabela 3.2 apresenta a avaliação do poder preditivo de cada variável em geral.

A variável **employed**, com um IV de 0,41, esta variável tem um poder preditivo forte. Isso indica que a situação de emprego é uma variável muito relevante. Com um IV de 0,33, a variável **region** também é um preditor forte. A região pode ser um fator significativo na análise de crédito. Com um IV de 0,64, a variável **age** é considerada um preditor muito forte. Isso sugere que a idade dos indivíduos é uma variável muito importante para prever a inadimplência. Com um IV de 0,42, o **marital_status** tem um poder preditivo forte. Isso indica que o estado civil é uma variável relevante. Com um IV de 0,02, a variável **gender** é considerada um preditor fraco. Isso sugere que o gênero dos indivíduos não é uma variável muito relevante para a análise.

Cálculos Realizados

1. **Cálculo das Proporções:** Cada variável é inicialmente convertida em uma proporção do total de participantes. Por exemplo, se 45% dos participantes são homens, a

proporção para `gender:male` é 0,45. Esse cálculo é feito para todas as variáveis e suas respectivas categorias.

2. **Cálculo de Eventos e Não Eventos:** Para cada categoria de variável, calculamos a proporção de eventos (inadimplência) e não eventos (não inadimplência). Isso nos ajuda a entender a distribuição de inadimplência dentro de cada categoria.

$$P(\textbf{event}) = \frac{\textbf{Número de inadimplentes na categoria}}{\textbf{Número total de inadimplentes}}$$

$$P(\textbf{non-event}) = \frac{\textbf{Número de não inadimplentes na categoria}}{\textbf{Número total de não inadimplentes}}$$

3. **Cálculo do Weight of Evidence (WoE):** O WoE é uma medida que compara a proporção de eventos e não eventos em uma categoria. Ele é calculado como o logaritmo natural da razão entre a proporção de eventos e a proporção de não eventos:

$$\text{WoE} = \ln\left(\frac{P(\text{event})}{P(\text{non-event})}\right)$$

4. **Cálculo do Information Value (IV):** O IV é uma medida da força da variável preditora. Ele é calculado como o produto da diferença entre as proporções de eventos e não eventos e o WoE:

$$\text{IV} = (P(\text{event}) - P(\text{non-event})) \times \text{WoE}$$

Foram analisadas cada uma das categorias das variáveis de origem e utiliza o **WoE** (Weight of Evidence) para analisar o peso de cada categoria na detecção da variável alvo.

Inicialmente, para cada categoria nas variáveis, foram criadas variáveis binárias para representar de forma simples se o indivíduo pertence ou não àquela categoria. Cada categoria representa uma característica do indivíduo.

As variáveis originais são irrelevantes para o Scorecard, já que cada categoria é analisada individualmente. Consequentemente, um indivíduo não pode pertencer a duas categorias da mesma variável.

São realizados novamente os cálculos de proporção, eventos e não eventos, WoE (Peso da Evidência) e IV (Valor da Informação) para cada uma das variáveis binárias.

Categoria	**%**	**Prop.**	**P_event**	**P_non_event**	**WoE**	**IV**
gender:male	0,45	675	0,20	0,24	-0,20	0,009
gender:female	0,55	825	0,30	0,24	0,20	0,011
employed:true	0,72	1080	0,51	0,20	0,94	0,29
employed:false	0,28	420	0,07	0,20	-0,94	0,11
age:18_30	0,27	405	0,07	0,19	-0,99	0,12
age:31_40	0,35	525	0,12	0,22	-0,61	0,06
age:41_50	0,23	345	0,05	0,17	-1,20	0,15
age:51_60	0,10	150	0,01	0,09	-2,19	0,17
age:60_100	0,05	75	0,002	0,04	-2,94	0,13
marital_status:married	0,54	810	0,29	0,24	0,16	0,006
marital_status:single	0,35	525	0,12	0,22	-0,61	0,064
marital_status:divorced	0,10	150	0,01	0,09	-2,19	0,17
marital_status:widower	0,10	150	0,01	0,09	-2,19	0,17
region:south_southeast	0,70	1050	0,49	0,21	0,84	0,23
region:north_northeast	0,30	450	0,09	0,21	-0,84	0,10

Tabela 3.3: Tabela de IV das variáveis

A tabela 3.3 representa os cálculos de **WoE** e **IV** de cada uma das categorias em cada variável.

3.3.3 Construção do Scorecard

O objetivo deste Scorecard é representar a aderência de um indivíduo ao perfil inadimplente conforme a pesquisa Boa Vista (2022). O Scorecard representa o peso da evidência mediante uma pontuação definida em uma escala. A cada variavel binária é atribuído uma pontuação. Esse Scorecard representa proporcionalmente a relação entre a maior pontuação e a propensão à inadimplência.

Cálculo dos Pontos: a fórmula utilizada para calcular a pontuação ajustados:

$$\text{Points} = \text{Factor} \times \text{WoE}$$

Onde o **Factor** é calculado como:

$$\text{Factor} = \frac{\text{PDO}}{\ln(2)}$$

O **PDO** (Points to Double the Odds) representa a mudança de pontos necessária para dobrar as chances de inadimplência.

6. **Ajuste da Pontuação:** Os pontos calculados são ajustados para se enquadrarem em uma escala definida (por exemplo, 2 a 10 pontos):

$$\text{Scaled Points} = \text{Min\_Score} + \left(\frac{\text{Points} - \min(\text{Points})}{\max(\text{Points}) - \min(\text{Points})} \right) \times (\text{Max\_Score} - \text{Min\_Score})$$

Essa normalização garante que todas as categorias se ajustem na faixa de pontuação definida.

Feature	Category	Score
gender	gender:male	8
gender	gender:female	8
employed	employed:true	8
employed	employed:false	6
age	age:18_30	6
age	age:31_40	7
age	age:41_50	6
age	age:51_60	4
age	age:60_100	2
marital_status	marital_status:married	8
marital_status	marital_status:single	7
marital_status	marital_status:divorced	4
marital_status	marital_status:widower	4
macroregion	macroregion:south_southeast	8
macroregion	macroregion:north_northeast	6

Tabela 3.4: Scorecard de Inadimplência

A tabela 3.4 representa a pontuação dada a cada categoria das variáveis utilizadas da pesquisa boa vista. A análise do Scorecard sugere que indivíduos financeiramente ativos, casados e residentes nas regiões mais desenvolvidas têm menor probabilidade de inadimplência. As variáveis demográficas e socioeconômicas desempenham um papel crucial na previsão da inadimplência, e as estratégias para mitigar esses riscos devem ser adaptadas a essas características. Para este Scorecard uma pontuação maior está relacionada a maior propensão a inadimplência.

3.4 Coleta de dados

De acordo com uma pesquisa recente conduzida pela Kepios inc. 2023. O Facebook tinha 2,989 mil milhões de utilizadores ativos mensais em abril de 2023, colocando-o em

primeiro lugar no ranking das plataformas de redes sociais mais “ativas” do mundo.

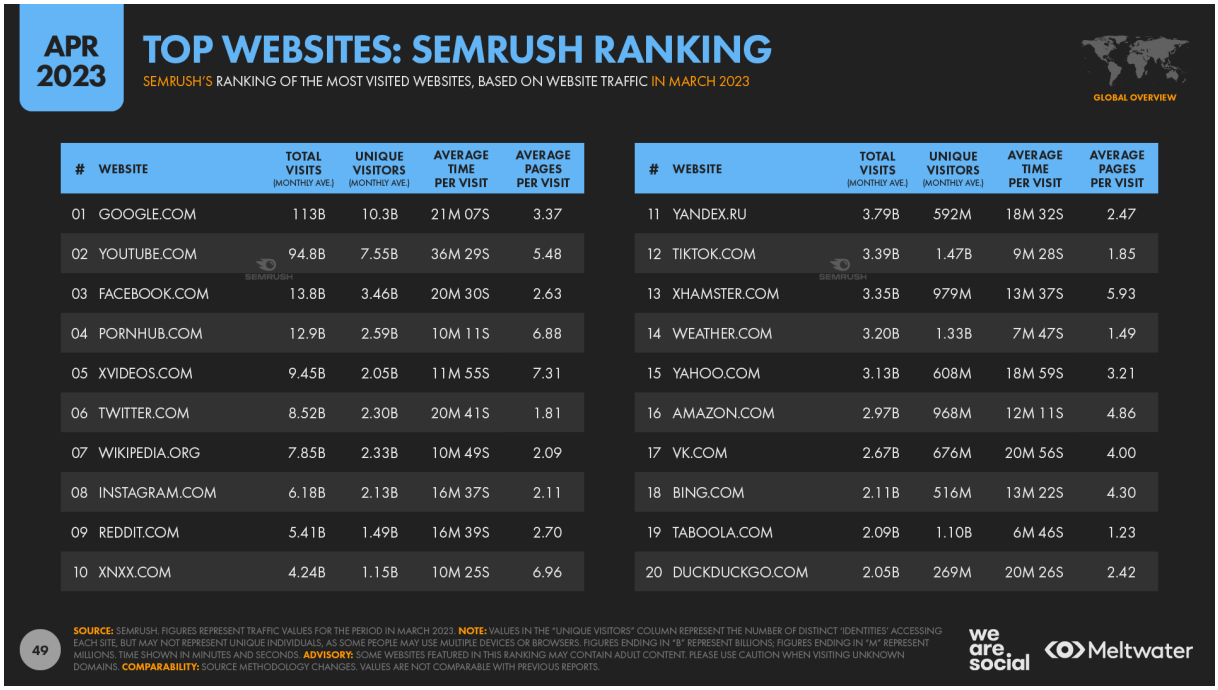

Figura 3.2: Rank Semrush para os sites mais acessados em 2023.

A figura 3.2 apresenta o ranking das redes sociais mais acessadas no mundo em 2023. Indicando o facebook como terceira rede social mais acessada.

A pesquisa da Kepios inc. 2023 indica que o número de usuários do Facebook no mundo ultrapassou 2,989 bilhões. 2,037 mil milhões utilizam o Facebook diariamente, o que significa que 68% das pessoas fazem login diariamente nas múltiplas plataformas, website e celulares. Os usuários ativos mensais do Facebook equivalem a 37,2% de todas as pessoas na Terra hoje. O Brasil tem em 2023 tem pelo menos 114,2 milhões de usuários ativos do Facebook.

Estes dados da Kepios inc. 2023 trazem provas suficientes da relevância dos dados e da presença do Facebook no cotidiano. O Facebook tem relevância para compreender as características das pessoas no Brasil e em todo o mundo. A coleta dos diferentes perfis de utilizadores do Facebook, foi conduzida uma coleta automatizada de dados pela técnica conhecida como Scraping. A extração permitiu a obtenção de informações detalhadas sobre os perfis dos usuários do Facebook, incluindo nome, idade, sexo, naturalidade, relacionamento, entre outros.

A coleta de dados iniciou-se em 2022. No processo de coleta, foi adotada uma abor-

dagem em duas etapas. Inicialmente, a busca foi iniciada a partir de um usuário central onde foram armazenados todos os dados do perfil e a coleta foi realizada, armazenando todas as interações públicas desse usuário e de seus amigos na rede social. Em seguida, a busca foi expandida para amigos dos amigos. Coletando todas as interações públicas do segundo nível da rede em relação ao usuário primário. O scraping abrangeu aproximadamente 130 mil perfis no Facebook. Isso permitiu a seleção e armazenamento de todas as interações públicas relevantes para a identificação de perfis de interação online.

O web scraping, apesar de ser uma prática amplamente utilizada, possui implicações legais que devem ser consideradas, especialmente no contexto brasileiro. Embora o Web scraping não seja necessariamente ilegal no Brasil, é fundamental adotar práticas éticas e responsáveis, respeitando os termos de serviço, evitando a sobrecarga de servidores e garantindo que os dados coletados sejam realmente de acesso público. O Facebook proíbe o Web scraping ou raspagem de dados de seus produtos e serviços sem permissão prévia aplicando medidas como detecção de padrões de atividade associados à automação de computadores e interrupção desses processos, envio de comunicados para cessar atividades de scraping. Entretanto, o Facebook reconhece que o scraping não pode ser totalmente evitado sem prejudicar o uso legítimo, com isso o facebook foca em dificultar a prática por meio de limites de taxa, limites de dados e outros métodos de detecção e prevenção Meta (2023). Nesse sentido, Franklin (2020b) ressalta a importância de respeitar os termos de serviço e evitar a raspagem excessiva de dados, a fim de não sobrecarregar os servidores. Além disso, é fundamental garantir que os dados coletados sejam realmente de acesso público e não violem direitos autorais ou a privacidade dos usuários.

Silva (2020) defende que o Web scraping em dados públicos, quando realizado de forma ética e responsável, pode ser uma ferramenta valiosa para a pesquisa e a análise de informações. No entanto, os autores ressaltam a necessidade de avaliar cuidadosamente os riscos legais e obter permissão prévia dos proprietários dos sites, quando necessário.

Variável	**Descrição,**	**Natureza**
id	O identificador único para o perfil.	Categórica
name	O nome completo do perfil.	Categórica
username	O nome de usuário associado ao perfil.	Categórica
gender	O sexo biológico do indivíduo.	Binária

Variável	**Descrição**	**Natureza**
email	O endereço de e-mail associado ao perfil.	Categórica
phone	O número de telefone associado ao perfil, se fornecido.	Categórica
birthday	A data de nascimento do perfil, formatada para substituir certos caracteres.	Categórica
age	A idade do perfil, se fornecida.	Contínua
languages	Idiomas falados pelo perfil.	Categórica
nick_name	Outros nomes ou apelidos associados ao perfil, formatados para remover certos caracteres.	Categórica
relationship	O estado de relacionamento do perfil, formatado para remover certos caracteres.	Categórica
about	A seção "Sobre Mim"do perfil, formatada para remover certos caracteres.	Categórica
religion	A afiliação religiosa do perfil, formatada para remover certos caracteres.	Categórica
political_statement	As preferências políticas do perfil, formatadas para remover certos caracteres.	Categórica
interests	Os interesses do perfil, formatados para remover certos caracteres.	Categórica
hometown	A cidade natal do perfil.	Categórica
hometown_state	O estado da cidade natal.	Categórica
hometown_state_code	O código do estado da cidade natal.	Categórica
hometown_state_region	A região do estado da cidade natal.	Categórica
nationality	A nacionalidade do perfil.	Categórica
current_city	A cidade de residência atual.	Categórica
current_state	O estado de residência atual.	Categórica
current_state_code	O código do estado de residência atual.	Categórica
current_state_region	A região do estado de residência atual.	Categórica
current_country	O país de residência atual.	Categórica

Variável	**Descrição**	**Natureza**
friend_count	O número de amigos que o perfil possui.	Discreta
follower_count	O número de seguidores que o perfil possui.	Discreta
following_count	O número de perfis que o usuário está seguindo.	Discreta
friends_link	Contagem de conexões abertas na rede social.	Discreta
jobs	Histórico de empregos cadastrados.	Categórica
jobs_count	Contagem de empregos registrados no perfil.	Discreta
education	O nível de educação do perfil.	Categórica

Tabela 3.5: Variáveis do perfil do Facebook

A tabela apresentada acima 3.5 lista as variáveis de perfil extraídas de dados do Facebook e descreve suas respectivas naturezas. Cada linha da tabela representa uma variável específica, com três colunas: “Nome da Variável”, “Descrição” e “Natureza da Variável”. A coluna “Nome da Variável” contém os nomes das variáveis usadas no conjunto de dados. A coluna “Descrição” fornece uma breve explicação sobre o que cada variável representa. A coluna “Natureza da Variável” categoriza a variável como categórica, contínua, discreta ou binária, dependendo de suas características. Esta classificação ajuda a entender melhor a natureza dos dados e a realizar análises apropriadas. Por exemplo, variáveis como ”id”e ”username”são categóricas, ao identificarem exclusivamente cada perfil, enquanto ”age”é uma variável contínua, representando a idade numérica do perfil.

3.4.1 Pré-Processamento

Os dados passaram por várias etapas de pré-processamento. Inicialmente os dados são extraídos cruamente em arquivos de texto. Foram removidas as marcações de página, símbolos, dados nulos, arquivos em branco. Os arquivos são formatados em dicionários Python e erros de formatação foram corrigidos. Os arquivos foram validados e salvos em JSON. Foram extraídos dados demográficos, nome, idade, data de nascimento, cidade natal e cidade de residência, lista de empregos, relacionamentos afetivos, entre outros. Os dados de amizades diretas estão também listados junto ao arquivo de perfil.

Após a formatação inicial, os dados de relacionamentos são extraídos e novos perfis foram adicionados a uma lista de scraping. De forma automática, todos os perfis dos listados foram extraídos com o processo de scraping. Novamente é repetido o trabalho de limpeza e formatação dos arquivos.

Novamente, os relacionamentos diretos são extraídos de cada um dos perfis já coletados do Facebook, a lista de novos perfis é então analisada para remover possíveis duplicações de perfis já coletados. Após a revisão da lista, os novos perfis foram coletados do Facebook.

3.4.2 Aplicação do Scorecard Boa Vista

Para aplicar o cálculo da pontuação do modelo Boa Vista no conjunto de dados do Facebook, primeiramente é necessário criar todas as variáveis equivalentes do Scorecard Boa Vista. Essas variáveis são preparadas para o cálculo antes da transformação. Categorias e variáveis fictícias são formatadas utilizando os valores do Scorecard. As variáveis são:

- **gender**: male, female
- **employed**: true, false
- **age**: 18-30, 31-40, 41-50, 51-60, 60-100
- **marital_status**: married, single, divorced, widower
- **region**: south_southeast, north_northeast

Essas variáveis foram criadas e separadas em um conjunto de dados e, em seguida, utilizadas para o cálculo da pontuação. Após o cálculo, o conjunto de dados retorna ao conjunto de dados original com a variável referente à pontuação para análise completa dos dados.

Criação da Variável gender

A variável gender indica o sexo biológico dos indivíduos. Durante o processo de limpeza e preparação dos dados, foi necessário padronizar os diversos valores presentes na base de dados para garantir a consistência e precisão na análise.

Inicialmente, diversos valores que indicavam formas variadas de masculino foram transformados para um valor padronizado masculino. Da mesma forma, valores que indicavam formas variadas de feminino foram transformados para um valor padronizado feminino. Valores de gênero não padronizados ou pouco claros foram definidos como NaN (Not a Number), indicando dados ausentes ou irrelevantes.

A padronização dos valores foi realizada conforme abaixo:

- **Masculino**: Valores como 'Homem', 'Gay', 'ALEGRE', 'Eu sou o Batman', 'FtM', 'Human', 'Masculino', 'Homossexual', 'Homossexual Passivo', 'Mulher (trans)', 'homo virilis, macho, versartil, incusubus e Exu Bombogiro', 'Trans homem', 'Homem (trans)', 'Transgênero', 'Epiceno', 'Lesbiana e Trans masculino', 'Tiste Andii', 'Not specified', 'Não Binário', 'Poliafrodita', 'Sagaz', 'Sem gênero', 'casado', 'curioso', 'genderqueer', 'homo', 'other', 'suspense', 'trans não binario' foram classificados como 'masculino'.
- **Feminino**: Valores como 'Cis Woman', 'Feminino', 'Trans mulher', 'Transgénero', 'Trans', 'Lesbiana e Trans masculino', 'Mulher (trans) e Mulher transexual', 'pleroma' foram classificados como 'feminino'.
- **NaN**: Valores como 'Masculino e Feminino', 'unisex', 'whatever', 'LGBT', 'Neutro', 'Ninguno', assexuada' foram definidos como NaN.

Valores classificados como NaN foram considerados irrelevantes para o modelo e, portanto, esses indivíduos foram removidos do conjunto de dados.

Após essa padronização, foi criada uma nova variável `d_gender` para codificar o gênero em uma variável numérica, onde 1 representa masculino e 0 representa feminino. Além disso, foram criadas as variáveis fictícias `gender:male` e `gender:female` desde a variável d_gender que serão utilizadas no cálculo da variável alvo com base na regressão logística. Essa transformação foi crucial para assegurar que a variável gênero pudesse ser corretamente interpretada e utilizada no modelo, permitindo uma análise mais precisa e coerente dos dados.

Criação da Variável Country

A variável país atual current_country indica o país de residência atual dos indivíduos.

O conjunto de dados foi filtrado para incluir apenas os indivíduos cujo país atual estava listado como 'Brasil' ou 'Brazil'. Todos os outros países foram excluídos da análise, uma vez que o foco do estudo era apenas nos indivíduos residentes no Brasil. Por fim, o nome do país foi padronizado para 'brazil' em todas as entradas, garantindo consistência na nomenclatura. Embora essa variável não seja usada no Scorecard, a remoção dos indivíduos de outros países é de suma importância para o cálculo correto do Scorecard. Essa transformação foi essencial para assegurar que apenas indivíduos vivendo atualmente no Brasil façam parte da pesquisa.

Criação Variável employed

A variável emprego employed indica a situação de emprego do indivíduo. Em equivalência com a variável employed da pesquisa Boa Vista que significa indivíduos empregados. Considerou-se economicamente ativo qualquer indivíduo que tenha um ou mais empregos registrados em seu perfil.

Foi definida uma função para converter a situação de emprego em uma variável binária. Se o número de empregos job_count for maior ou igual a 1, a função retorna 1 empregado; caso contrário, retorna 0 não empregado. Para a modelagem, a variável de emprego binário foi utilizado para criar variáveis fictícias: employed:true e employed:false.

Criação da Variável age

A variável idade (*age*) indica a idade dos indivíduos. Indivíduos com idades entre 16 e 18 anos foram redefinidos para 18 anos para os fins desta pesquisa. Todos os indivíduos abaixo de 16 anos foram removidos do conjunto de dados. Os dados para indivíduos com 90 anos ou mais foram revisados e removidos do modelo. As idades foram agrupadas nas seguintes categorias, considerando o Scorecard da pesquisa Boa Vista: 18-30, 31-40, 41-50, 51-60, 60-100. Para a modelagem, foram criadas variáveis fictícias para cada grupo etário. age:18-30, age:31-40, age:41-50, age:51-60 e age:60-100.

Criação da Variável region

A variável região (*region*) indica a região de residência dos indivíduos. Foi definida uma função para converter os nomes das regiões em categorias numéricas conforme a classificação baseada na pesquisa Boa Vista. A função foi aplicada à coluna current_state_region, para criar uma nova coluna numérica region. Para a modelagem, foram

criadas as variáveis fictícias para cada região region:south, region:southeast, region:north, region:northeast, region:midwest.

Criação da Variável marital_status

A variável estado civil marital_status indica o estado civil dos indivíduos. Primeiramente, uma função foi aplicada à coluna relationship para criar uma nova coluna marital_status. A função classifica o estado civil dos indivíduos. Para a modelagem, foram criadas variáveis fictícias para cada categoria de estado civil: marital_status:married, marital_status:single, marital_status:divorced, marital_status:widower

3.4.3 Calculo da variável proba

Conforme o modelo criado anteriormente. Foi separado o mesmo conjunto de variáveis. Após isso foi aplicada a multiplicação dos dados da pontuação de cada variável do Scorecard com o valor da classe do conjunto extraído dos dados de rede social. Os valores desta função foram atribuídos a variável proba determinado a pontuação da probabilidade do indivíduo ser inadimplente.

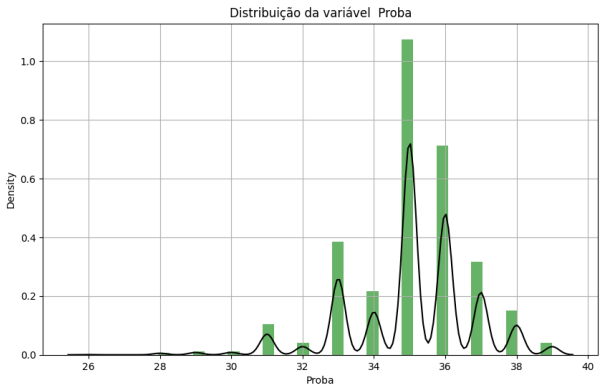

Figura 3.3: Distribuição da variável proba

O gráfico 3.3 representa a distribuição dos valores da variável proba.

3.4.4 Calculo da variável default

A variável default armazena as informações de se o indivíduo tem uma alta probabilidade de inadimplência durante um longo período. Conforme o Scorecard. A variável

default representa indivíduos com chance de deixar de pagar a dívida mesmo após tentativas de negociação amigável. Onde a negativação do devedor deve ser efetuada. A variável default é calculada utilizando a Análise Discriminante Linear (LDA) para encontrar o ponto de corte que represente da melhor forma a separação entre as classes da Variável default. Em busca do cálculo do valor de corte foram necessárias a criação de variáveis de suporte ao cálculo.

Utilizando os dados da Variável proba, foi criada uma Variável fictícia chamada **event** que representa a Variável proba de forma binaria dividida pela mediana de proba.

$$\text{event}(x_i) = \begin{cases} 1 & \text{se } x_i \geq \text{median}(x) \\ 0 & \text{se } x_i < \text{median}(x) \end{cases}$$

onde x_i representa os valores individuais da variável `proba` e median(x) representa a mediana dos valores de `proba`. A variável proba foi então dividida em 10 intervalos e armazenada na variável proba_bin para facilitar o cálculo do WoE. O WoE foi então calculado utilizando os bins da variável proba e a variável **event** como objetivo da função. os valores de WoE foram então armazenados na coluna WoE. O ajuste do modelo de análise discriminatória LDA é então realizado utilizando os valores de WoE e a variável **event**. A pontuação LDA é calculada e o ponto de corte é definido como a média dessas pontuações. A variável `default` é criada com base no ponto de corte das pontuações LDA:

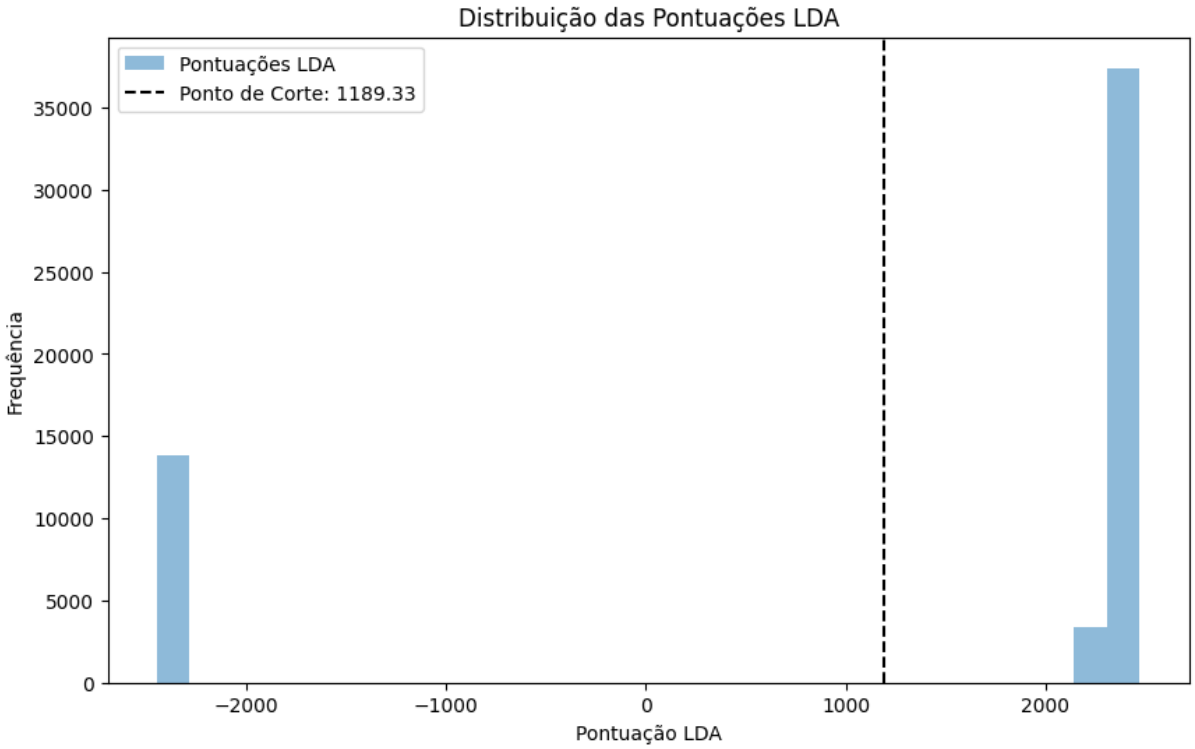

Figura 3.4: Distribuição das pontuações LDA

O ponto de corte é a média das pontuações LDA e divide a distribuição em duas partes. A escolha deste ponto de corte é crucial para a classificação, ao determinar quais observações são consideradas inadimplentes ou adimplentes. Embora não se possa visualizar diretamente a separação das classes neste gráfico, a simetria sugere que o ponto de corte é uma boa escolha para dividir as observações em dois grupos distintos. O gráfico 3.4 permite visualizar como a probabilidade de inadimplência (proba) se distribui entre as duas classes. O gráfico mostra uma distribuição bimodal, onde podemos ver duas curvas distintas, uma representando a classe adimplente e outra a classe inadimplente. Isso indica que existe uma separação razoável entre as duas classes com base na variável proba. A linha pontilhada vertical indica o ponto de corte (proba_cutoff). Este é o valor de proba correspondente ao ponto de corte calculado com base na pontuação LDA. A presença desta linha facilita a visualização de como os dados são divididos em adimplentes e inadimplentes. Existe alguma sobreposição entre as duas distribuições, o que sugere que alguns adimplentes possuem probabilidade de inadimplência similar a alguns inadimplentes. Esta sobreposição é um desafio comum em modelos de classificação e indica a necessidade de um ponto de corte bem definido.

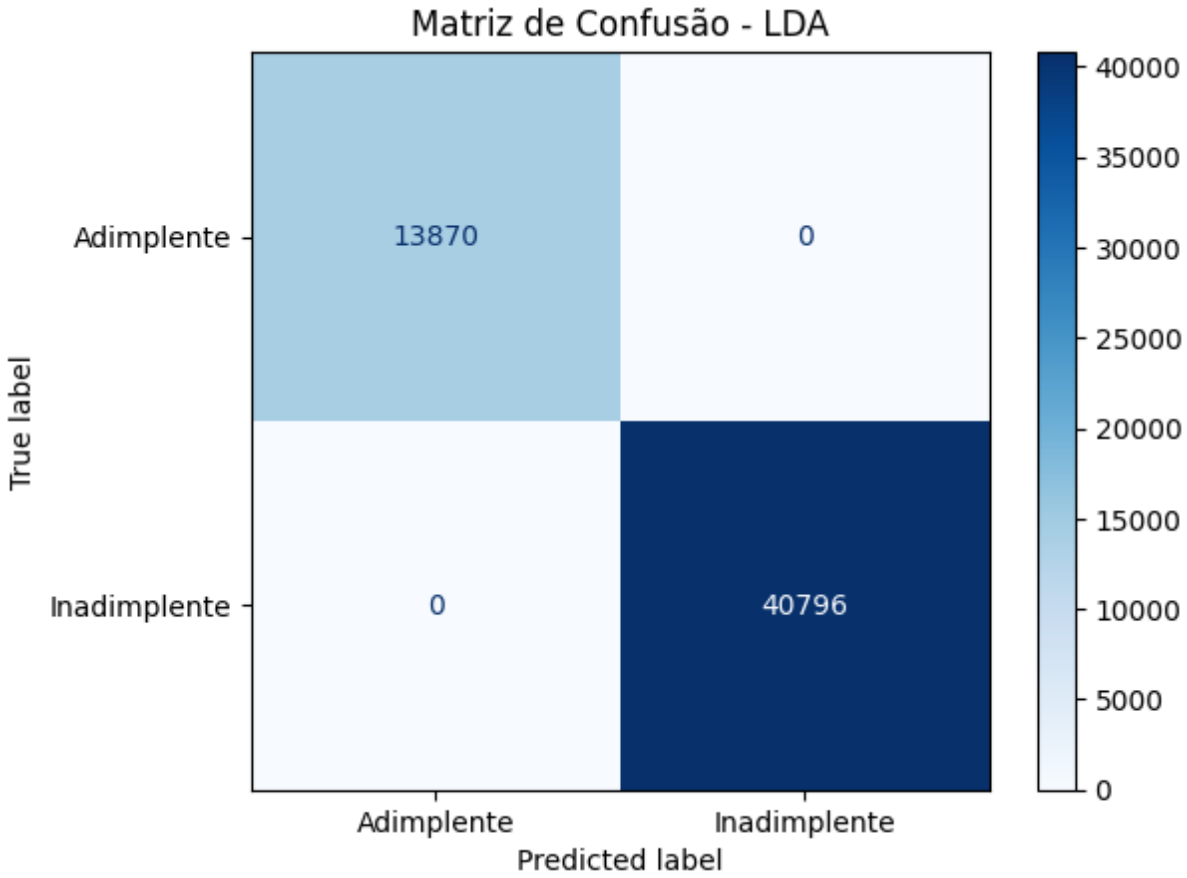

Figura 3.5: Matriz de confusão LDA

A matriz de confusão 3.5 compara as previsões de default feitas pelo modelo com as classes reais, apresentando a capacidade do modelo em separar ambas as classes.

Os gráficos mostram que o modelo LDA é capaz de fornecer uma separação razoável entre as classes adimplente e inadimplente com base nas pontuações calculadas. A escolha do ponto de corte parece adequada, como indicado pela distribuição das pontuações LDA. No entanto, a sobreposição observada na distribuição de proba sugere haver espaço para melhorar a discriminação entre as classes. A matriz de confusão fornece uma visão detalhada do desempenho do modelo e ajuda a identificar áreas onde o modelo pode ser aprimorado.

3.4.5 Análise exploratória

O conjunto de dados extraído do Facebook apresenta uma abundância de novas variáveis à análise de crédito, oferecendo uma perspectiva única e abrangente sobre o com-

portamento e as características dos usuários. Essas variáveis podem incluir informações demográficas, interesses, atividades, redes de contatos e outros dados comportamentais que podem ser cruciais para a avaliação de crédito.

A análise destas variáveis visou entender a relevância de cada variável no contexto da análise de crédito. Buscou-se compreender se existem variáveis que apresentem forte correlação com a inadimplência e quais variáveis são pouco relevantes ou irrelevantes. identificou-se variáveis que podem introduzir vieses ou discriminação no modelo de crédito. Variáveis que, direta ou indiretamente, discriminam com base em raça, gênero, idade, ou outros atributos protegidos, devem ser removidas para garantir a equidade e a conformidade com regulamentações legais e éticas. A análise visou entender a relevância de cada variável para a predição da inadimplência, através do cálculo do peso da evidência e valor da informação de cada variável é possível identificar quais são relevantes, irrelevantes e detectar valores atípicos. Ao valor de cada variável foi analisado o IV e classificada como preditor irrelevante, baixo, médio, forte e muito forte, também considerado suspeito. as categorias são então ajustadas utilizando técnicas de engenharia para garantir que os dados sejam representados da melhor forma. A análise da variável inclui a análise dos valores, preenchimento de linhas em branco, cálculos de WoE e IV e transformação das variáveis. A análise de cada variável utiliza os seguintes cálculos:

- **variavel**: categorias da variável.
- **n_obs**: número de observações.
- **prop_bom**: proporção de bons pagadores.
- **prop_n_obs**: proporção de observações.
- **n_bom**: número de bons pagadores.
- **n_ruim**: número de maus pagadores.
- **prop_n_bom**: proporção de bons pagadores no total.
- **prop_n_ruim**: proporção de maus pagadores no total.
- **WoE**: Weight of Evidence.

- **diff_prop_bom**: diferença na proporção de bons pagadores.
- **diff_WoE**: Diferença no Weight of Evidence.
- **IV**: Information Value.

Para a análise do IV information value o valor foi classificado conforme a tabela abaixo 3.6.

Faixa de IV	**Classificação**
0,00 - 0,02	Baixo
0,02 - 0,1	Médio
0,1 - 0,3	Bom
0,3 - 0,5	Forte
$> 0{,}5$	Suspeito

Tabela 3.6: Classificação dos valores de Information Value (IV).

Análise da variável about A variável **about** armazena uma breve descrição sobre o indivíduo. Ao analisar a variável foi detectado que 90% dos indivíduos não contem a informação, transformando-a em irrelevante para o estudo. Sendo removida do conjunto de dados.

Análise da variável age A variavel age classifica a idade dos indivíduos conforme o Scorecard da pesquisa boa vista.

age	n_obs	n_good	n_bad	WoE	IV
18-30	2517	0,05	2517,05	-11,90	2,16
31-40	3896	434,05	3462,05	-3,15	0,75
41-50	23047	18718,05	4329,05	0,38	0,05
51-60	7981	6668,05	1313,05	0,54	0,03
60-100	17151	14921,05	2230,05	0,82	0,16

Tabela 3.7: Tabela de análise por idade

Conforme a tabela 3.7Os dados analisados permitem identificar que a faixa etária "41-50"tem o maior número de observações (23047), enquanto a faixa etária "18-30"tem

o menor número de observações (2517). Além disso, a faixa etária "18-30"apresenta a menor proporção de bons pagadores e a faixa etária "60-100"tem a maior proporção de bons pagadores. A maior proporção de observações é da faixa "41-50", e a menor proporção de observações é da faixa "18-30". No total, a faixa "41-50"tem a maior proporção de bons pagadores, enquanto a faixa "18-30"tem a menor proporção de bons pagadores A faixa "31-40"(classificação 1) tem a maior proporção de maus pagadores no total, e a faixa "60-100"tem a menor proporção de maus pagadores no total (0,09). A faixa "18-30"apresenta o menor valor de WoE (-11,90), indicando um risco considerável, enquanto a faixa "60-100"apresenta o maior valor de WoE (0,82), indicando menor risco. O IV mais alto é encontrado na faixa "18-30"(2,16), indicando alta discriminação de risco, enquanto o IV mais baixo é na faixa "51-60"(0,03), indicando baixa discriminação de risco. A faixa etária "18-30"possui um alto risco de inadimplência, conforme indicado pelo alto IV e baixo WoE. As faixas etárias "41-50"e "60-100"mostram bons sinais de crédito, com altos valores de prop_bom e baixos valores de prop_n_ruim. A faixa "31-40"também apresenta um risco considerável, com uma alta proporção de maus pagadores (prop_n_ruim). A análise da tabela indica que indivíduos entre 60 e 100 anos têm um valor de informação suspeito para a avaliação do modelo.

Análise da variável age_group

Com os dados da idade dos indivíduos foi possível reorganizar a informação de idade dos indivíduos em menores intervalos onde o valor da informação esteja melhor representada. A idade dos indivíduos foi separada em intervalos conforme o IV para remover valores suspeitos.

Value	All	ruim	bom	WoE	IV
36,00 - 39,00	3856	3324	532	-0,75	0,03
56,00 - 117,00	3750	205	3545	3,92	0,98
34,00 - 36,00	3265	2807	458	-0,73	0,02
26,00 - 28,00	5813	4782	1031	-0,45	0,01
17,00 - 23,00	4939	3803	1136	-0,12	0,001
31,00 - 32,00	2497	2186	311	-0,87	0,02
29,00 - 31,00	4915	4266	649	-0,80	0,04
28,00 - 29,00	2782	2322	460	-0,54	0,01
32,00 - 34,00	4001	3502	499	-0,86	0,04
39,00 - 43,00	4355	3693	662	-0,64	0,02
23,00 - 25,00	4434	3594	840	-0,37	0,01
43,00 - 48,00	3572	2981	591	-0,53	0,016
25,00 - 26,00	2506	2022	484	-0,35	0,005
48,00 - 56,00	3907	1254	2653	1,82	0,29

Tabela 3.8: Análise da variável age_group.

A tabela 3.8 representa os novos intervalos de idade onde é possível identificar no valor do information value das categorias que foram ajustados os valores suspeitos.

Análise da variável marital_status

A variável marital_status representa o estado civil dos indivíduos no conjunto de dados e inclui quatro categorias: casado[0], solteiro[1], divorciado[2] e viúvo[3].

marital_status	n_obs	n_bom	n_ruim	WoE	IV
2	471	10,05	461,05	-4,90	0,16
3	223	8,05	215,05	-4,36	0,06
1	42920	31996,05	10924,05	-0,004	0,01
0	10978	8727,05	2251,05	0,27	0,014

Tabela 3.9: Análise da variável marital_status

A análise desta variável revela algumas diferenças significativas nas proporções de

boas e más observações entre as categorias. Conforme a tabela 3.9 a categoria de divorciados apresenta uma baixa proporção de boas observações (2,12%) e um WoE (Weight of Evidence) significativamente negativo (-4,90), indicando uma forte associação com más observações. Já a categoria de viúvos, também mostra uma baixa proporção de boas observações (3,59%) e um WoE negativo (-4,36), mas um pouco melhor que a dos divorciados. O Information Value (IV) de ambas as categorias é relativamente baixo, sugerindo menor importância preditiva. Por outro lado, a categoria de solteiros é a mais representada no conjunto de dados, com 78,62% das observações, e possui uma alta proporção de boas observações (74,55%). No entanto, o WoE próximo de zero (-0,004) indica que esta categoria não contribui significativamente para a discriminação entre boas e más observações, o que é reforçado pelo IV muito baixo.

A categoria de casados tem a maior proporção de boas observações (79,50%) e um WoE positivo (0,27), sugerindo uma associação favorável com boas observações. O Information Value (IV) desta categoria indica uma importância moderada na predição. Em suma, as categorias de divorciados e viúvos têm associações negativas com boas observações, enquanto a categoria de solteiros, embora prevalente, não discrimina bem entre boas e más observações. A categoria de casados, por sua vez, apresenta a melhor associação com boas observações. O Information Value total é baixo, indicando que a variável marital_status tem uma importância moderada na modelagem preditiva.

Análise da variável employed

A variável employed apresenta dois valores distintos, 0 e 1, que representam respectivamente indivíduos desempregados e empregados.

employed	n_obs	n_good	n_bad	WoE	IV
0	46509	33100,05	13409,05	-0,17	0,027
1	8083	7641,05	442,05	1,77	0,27

Tabela 3.10: Análise da variavel employed

Conforme a tabela acima 3.10 Para os desempregados, temos 46,509 observações, representando 85,19% do total de observações. A proporção de bons pagadores prop_bom é de 71,17%. O Weight of Evidence WoE é -0,17, indicando uma menor propensão ao

bom crédito comparado aos empregados. O valor de IV é 0,02, sugerindo uma variável de baixa importância preditiva.

Para os empregados ($employed = 1$), temos 8,083 observações, representando 14,81% do total de observações. A proporção de bons pagadores (*prop_bom*) é de 94,53%. A proporção de bons pagadores (*prop_n_bom*) é 18,76%, e a proporção de maus pagadores (*prop_n_ruim*) é 3,19%. O WoE é 1,77, indicando uma maior propensão ao bom crédito comparado aos desempregados. O valor de IV é 0,27, sugerindo uma variável de média importância preditiva.

A análise do WoE e do IV (Information Value) indica que a condição de emprego é um fator muito significativo para a probabilidade de bom crédito, sendo mais preditivo para indivíduos empregados do que desempregados. A significativa prevalência de desempregados cria uma distorção capaz de gerar um overfitting no modelo.

Análise da variável job_count

job_count	n_obs	n_good	n_bad	WoE	IV
0	46509	33100,05	13409,05	-0,17	0,02
1	7667	7244,05	423,05	1,76	0,25
3	83	79,05	4,05	1,89	0,003
2	333	318,05	15,05	1,97	0,01

Tabela 3.11: Análise da variável job_count

Conforme os dados da tabela 3.11. A categoria 0 (sem emprego) representa a maioria significativa das observações, com 46509 casos (85,19% do total). A categoria 1 (um emprego) é a segunda mais frequente, com 7667 observações (14,04%). As categorias 2 e 3 são consideravelmente menos frequentes. Todas as categorias apresentam uma alta proporção de casos bons, porém com diferenças notáveis. A categoria 0 tem a menor proporção de casos bons (71,17%), enquanto as categorias 1, 2 e 3 apresentam proporções significativamente mais altas, todas acima de 94%. A categoria 0 é a única com WoE negativo (-0,17), indicando uma associação negativa ligeiramente com casos bons. As categorias 1, 2 e 3 apresentam WoE positivo e elevado, com a categoria 2 tendo o maior valor (1,97). Isso sugere que ter pelo menos um emprego está fortemente associado a um resultado positivo em termos de risco de crédito.O IV total da variável é 0,303065,

indicando um poder preditivo médio.

Análise da variável Gender

A variável *gender* apresenta dois valores distintos, 0 e 1, que representam respectivamente indivíduos do gênero feminino e masculino.

gender	n_obs	n_bom	n_ruim	WoE	IV
0	21298	15294,05	6004,05	-0,14	0,008
1	33294	25447,05	7847,05	0,09	0,005

Tabela 3.12: Análise da variável gender

A tabela acima 3.12 mostra que para as mulheres ($gender = 0$), temos 21298 observações, representando 39,01% do total de observações. A proporção de boas pagadoras (*prop_bom*) é de 71,81%. O número de boas pagadoras (*n_bom*) é 15294,05, enquanto o número de más pagadoras (*n_ruim*) é 6004,05. A proporção de boas pagadoras (*prop_n_bom*) é 37,54%, e a proporção de más pagadoras (*prop_n_ruim*) é 43,35%. O Weight of Evidence (WoE) é -0,14, indicando uma menor propensão ao bom crédito comparado aos homens. O valor de IV (Information Value) é 0,008, sugerindo uma variável de baixa importância preditiva. Para os homens ($gender = 1$), temos 33294 observações, representando 60,99% do total de observações. A proporção de bons pagadores (*prop_bom*) é de 76,43%. O número de bons pagadores (*n_bom*) é 25447,05, enquanto o número de maus pagadores (*n_ruim*) é 7847,05. A proporção de bons pagadores (*prop_n_bom*) é 62,46%, e a proporção de maus pagadores (*prop_n_ruim*) é 56,65%. O WoE é 0,09, indicando uma maior propensão ao bom crédito comparado às mulheres. O valor de IV é 0,005, sugerindo uma variável de baixa importância preditiva.

A análise do WoE (Weight of Evidence) e do IV (Information Value) indica que o gênero possui uma diferença na probabilidade de bom crédito, sendo ligeiramente mais preditivo para homens do que para mulheres, embora sua importância preditiva geral seja baixa.

Análise da variável birthday

A variável birthday representa a data de aniversário dos indivíduos, através dela foi possível calcular a idade dos indivíduos na variável age. A variável contem uma alta

quantidade de categorias de difícil processamento, por tanto foi removida do conjunto de dados para processamento final.

Análise da variável current_City

A variavel current_city representa a cidade onde o individuo reside.Ao analisar os dados da variavel, é possível entender diretamente que se trata de uma variavel com distintos valores onde ao classificar em categorias o valor da informação IV é redundante. Com isto a variavel não tem valor estatístico para o modelo.

Análise da variável is_capital_city

Através da análise da cidade de residência do indivíduo e dados do IBGE, foi possível construir uma variável indicadora para verificar se ele reside em uma capital ou estado. Com isso, foi criada a variável iscapitalcity. Os novos dados gerados foram armazenados na variável current_city_encoded. A variável is_capitalcity apresenta dois valores distintos, 0 e 1, que representam respectivamente indivíduos que não vivem e que vivem em uma capital ou região metropolitana dos estados.

is_capitalcity	**n_obs**	**n_bom**	**n_ruim**	**WoE**	**IV**
0	4440	2921,05	1519,05	-0,42	0,01
1	50152	37820,05	12332,05	0,04	0,001

Tabela 3.13: Análise da variável is_capitalcity

Conforme a análise da tabela acima 3.13. Para os indivíduos que vivem em uma capital ou região metropolitana, temos 4440 observações, representando 8,13% do total de observações. A proporção de bons pagadores é de 65,79%. O número de bons pagadores é 2921,05, enquanto o número de maus pagadores é 1519,05. O Weight of Evidence é -0,42, indicando uma menor propensão ao bom crédito comparado aos que não vivem em uma capital ou região metropolitana. O valor de é 0,01, sugerindo uma variável de baixa importância preditiva.

Para os indivíduos que não vivem em uma capital ou região metropolitana temos 50152 observações, representando 91,87% do total de observações. A proporção de bons pagadores é de 75,41%. O número de bons pagadores é 37820,05, enquanto o número de maus pagadores é 12332,05. A proporção de bons pagadores é 92,83%, e a proporção

de maus pagadores é 89,03%. O WoE é 0,04, indicando uma maior propensão ao bom crédito comparado aos que vivem em uma capital ou região metropolitana. O valor de IV é 0,001, sugerindo uma variável de baixa importância preditiva.

A análise do WoE e do IV indica que viver ou não em uma capital ou região metropolitana possui uma diferença na probabilidade de bom crédito, sendo ligeiramente mais preditivo para aqueles que não vivem em uma capital ou região metropolitana, embora sua importância preditiva geral seja baixa.

Análise da variável current_state

current_state	**n_obs**	**n_bom**	**n_ruim**	**WoE**	**IV**
roraima	36	4,05	32,05	-3,14	0,006
sergipe	138	16,05	122,05	-3,10	0,02
piaui	601	85,05	516,05	-2,88	0,10
maranhao	963	142,05	821,05	-2,83	0,15
amazonas	293	44,05	249,05	-2,81	0,04
bahia	1077	174,05	903,05	-2,72	0,16
tocantins	123	20,05	103,05	-2,71	0,01
alagoas	598	99,05	499,05	-2,69	0,09
amapa	42	7,05	35,05	-2,68	0,006
distrito_federal	282	53,05	229,05	-2,54	0,03
rondonia	234	45,05	189,05	-2,51	0,03
paraiba	506	99,05	407,05	-2,49	0,06
rio_grande_do_norte	504	102,05	402,05	-2,44	0,06
goias	468	100,05	368,05	-2,38	0,05
ceara	1552	350,05	1202,05	-2,31	0,18
para	473	113,05	360,05	-2,23	0,05
pernambuco	2379	570,05	1809,05	-2,23	0,26
mato_grosso	186	45,05	141,05	-2,22	0,02
mato_grosso_do_sul	93	23,05	70,05	-2,19	0,009
acre	34	9,05	25,05	-2,09	0,003
espirito_santo	541	406,05	135,05	0,02	0,001
rio_grande_do_sul	660	561,05	99,05	0,65	0,004

current_state	n_obs	n_bom	n_ruim	WoE	IV
minas_gerais	18465	15719,05	2746,05	0,66	0,12
rio_de_janeiro	4428	3858,05	570,05	0,83	0,04
santa_catarina	887	796,05	91,05	1,08	0,01
parana	1055	947,05	108,05	1,09	0,01
sao_paulo	17974	16354,05	1620,05	1,23	0,35

Tabela 3.14: Análise da variável current_state

Analisando os dados conforme a tabela acima 3.14. A variável current_state representa os estados brasileiros dos indivíduos. A tabela acima mostra várias métricas relacionadas à distribuição e qualidade de crédito dos indivíduos de diferentes estados.

Roraima apresenta uma proporção de bons pagadores de 11,11% com um Weight of Evidence WoE de -3,14, indicando uma propensão muito baixa ao bom crédito. Sergipe também apresenta uma baixa proporção de bons pagadores (11,59%) e um WoE de -3,10. Estados como Piauí, Maranhão, Amazonas e Bahia também mostram uma propensão negativa ao bom crédito, com WoE variando de -2,88 a -2,72 .

Por outro lado, estados como Espírito Santo (75,05%), Rio Grande do Sul (85%), Minas Gerais (85,13%), Rio de Janeiro (87,13%), Santa Catarina (89,74%), Paraná (89,76%) e São Paulo (90,99%) mostram uma alta proporção de bons pagadores, com WoE variando de 0,02 a 1,23. Isso indica uma maior propensão ao bom crédito comparado aos estados com proporções mais baixas.

A análise do *WoE* (Weight of Evidence) e do *IV* (Information Value) indica que o estado atual de residência é um fator significativo para a probabilidade de bom crédito. Estados com proporções mais altas de bons pagadores e WoE bons são indicativos de melhores comportamentos de crédito.

Análise da variável hometown_state_region

A variável *hometown_state_region* representa a região de origem dos indivíduos.

hometown_state_region	n_obs	n_bom	n_ruim	WoE	IV
norte	999	317,05	682,05	-1,84	0,07
nordeste	7483	2717,05	4766,05	-1,64	0,45
centro-oeste	775	309,05	466,05	-1,48	0,03
sudeste	43510	35883,05	7627,05	0,46	0,15
sul	1825	1515,05	310,05	0,50	0,007

Tabela 3.15: Análise da variável hometown_state_region

A análise da informação da tabela acima 3.15. Para a região Norte, temos 999 observações, representando 1,83% do total de observações. A proporção de bons pagadores é de 31,73%, com um número de bons pagadores de 317,05 e maus pagadores de 682,05. O WoE é -1,84, indicando uma menor propensão ao bom crédito. O valor de IV é 0,07, sugerindo uma variável de baixa a moderada importância preditiva.

A região Nordeste possui 7483 observações, representando 13,71% do total. A proporção de bons pagadores é de 36,31%, com um WoE de -1,64, também indicando uma menor propensão ao bom crédito. O IV é 0,45, sugerindo uma importância preditiva moderada.

A região Centro-Oeste tem 775 observações, representando 1,42% do total, com uma proporção de bons pagadores de 39,87%. O WoE é -1,48, e o IV é 0,038, indicando uma variável de baixa importância preditiva.

Para a região Sudeste, temos 43510 observações, representando 79,70% do total. A proporção de bons pagadores é de 82,47%, com um WoE de 0,46, indicando uma maior propensão ao bom crédito. O IV é 0,15, sugerindo uma importância preditiva moderada.

A região Sul possui 1825 observações, representando 3,34% do total. A proporção de bons pagadores é de 83,01%, com um WoE de 0,50, também indicando uma maior propensão ao bom crédito. O IV é 0,007, sugerindo uma variável de baixa importância preditiva.

A análise do WoE e do IV indica que a região de origem dos indivíduos é um fator significativo para a probabilidade de bom crédito, com as regiões Sudeste e Sul apresentando maior propensão ao bom crédito em comparação às regiões Norte, Nordeste e

Centro-Oeste.

Análise da variável Current_State_Region

A variável *current_state_region* representa a região de residência atual dos indivíduos.

current_state_region	n_obs	n_bom	n_ruim	WoE	IV
norte	999	317,05	682,05	-1,84	0,07
nordeste	7483	2717,05	4766,05	-1,64	0,45
centro-oeste	775	309,05	466,05	-1,48	0,03
sudeste	43510	35883,05	7627,05	0,46	0,15
sul	1825	1515,05	310,05	0,50	0,007

Tabela 3.16: Análise da variável current_state_region

A tabela 3.16 mostra várias métricas relacionadas à distribuição e qualidade de crédito dos indivíduos de diferentes regiões.

Para a região Norte, temos 999 observações, representando 1,83% do total de observações. A proporção de bons pagadores (*prop_bom*) é de 31,73%, com um número de bons pagadores (*n_bom*) de 317,05 e maus pagadores (*n_ruim*) de 682,05. O Weight of Evidence (WoE) é -1,84, indicando uma menor propensão ao bom crédito. O valor de IV (Information Value) é 0,07, sugerindo uma variável de baixa a moderada importância preditiva.

A região Nordeste possui 7483 observações, representando 13,71% do total. A proporção de bons pagadores é de 36,31%, com um WoE de -1,64, também indicando uma menor propensão ao bom crédito. O IV é 0,45, sugerindo uma importância preditiva moderada.

A região Centro-Oeste tem 775 observações, representando 1,42% do total, com uma proporção de bons pagadores de 39,87%. O WoE é -1,48, e o IV é 0,03, indicando uma variável de baixa importância preditiva.

Para a região Sudeste, temos 43510 observações, representando 79,70% do total. A proporção de bons pagadores é de 82,47%, com um WoE de 0,46, indicando uma maior propensão ao bom crédito. O IV é 0,15, sugerindo uma importância preditiva moderada.

A região Sul possui 1825 observações, representando 3,34% do total. A proporção de bons pagadores é de 83,01%, com um WoE de 0,50, também indicando uma maior propensão ao bom crédito. O IV é 0,007, sugerindo uma variável de baixa importância preditiva.

A análise do WoE (Weight of Evidence) e do IV (Information Value) indica que a região de residência atual dos indivíduos é um fator significativo para a probabilidade de bom crédito, com as regiões Sudeste e Sul apresentando maior propensão ao bom crédito em comparação às regiões Norte, Nordeste e Centro-Oeste.

Análise da variável region

A variável *region* representa diferentes regiões onde os indivíduos residem. A tabela 3.17 mostra várias métricas relacionadas à distribuição e qualidade de crédito dos indivíduos de diferentes regiões.

Tabela 3.17: Análise da variável region

region	n_obs	n_bom	n_ruim	WoE	IV
2	1150	165,05	985,05	-2,86	0,19
3	7909	1254,05	6655,05	-2,74	1,23
4	993	187,05	806,05	-2,53	0,13
1	40523	35554,05	4969,05	0,88	0,45
0	4017	3581,05	436,05	1,02	0,05

Para a região 2, temos 1150 observações, representando 2,11% do total de observações. A proporção de bons pagadores (*prop_bom*) é de 14,35%, com um número de bons pagadores (*n_bom*) de 165,05 e maus pagadores (*n_ruim*) de 985,05. O Weight of Evidence (WoE) é -2,86, indicando uma menor propensão ao bom crédito. O valor de IV (Information Value) é 0,19, sugerindo uma variável de moderada importância preditiva.

A região 3 possui 7909 observações, representando 14,49% do total. A proporção de bons pagadores é de 15,86%, com um WoE de -2,74, também indicando uma menor propensão ao bom crédito. O IV é 1,23, sugerindo uma importância preditiva alta.

A região 4 tem 993 observações, representando 1,82% do total, com uma proporção de bons pagadores de 18,83%. O WoE é -2,53, e o IV é 0,13, indicando uma variável de

moderada importância preditiva.

Para a região 1, temos 40523 observações, representando 74,23% do total. A proporção de bons pagadores é de 87,74%, com um WoE de 0,88, indicando uma maior propensão ao bom crédito. O IV é 0,45, sugerindo uma importância preditiva alta.

A região 0 possui 4017 observações, representando 7,36% do total. A proporção de bons pagadores é de 89,15%, com um WoE de 1,02, também indicando uma maior propensão ao bom crédito. O IV é 0,05, sugerindo uma variável de baixa importância preditiva.

A análise do WoE (Weight of Evidence) e do IV (Information Value) indica que a região de residência dos indivíduos é um fator significativo para a probabilidade de bom crédito, com as regiões 1 e 0 apresentando maior propensão ao bom crédito em comparação às regiões 2, 3 e 4.

Transformação da Variável languages

Foi transformada para definir a quantidade de idiomas que cada indivíduo fala e, após isso, categorizada em: 1 idioma, bilíngue, trilíngue e poliglota.

Análise da variável language_count

A variável language_count representa a quantidade de idiomas que um indivíduo fala. A tabela 3.18 mostra várias métricas relacionadas à distribuição e qualidade de crédito dos indivíduos com diferentes quantidades de idiomas falados.

language_count	n_obs	n_bom	n_ruim	WoE	IV
10	1	0,05	1,05	-4,12	0,0003
48	1	0,05	1,05	-4,12	0,000308
6	6	4,05	2,05	-0,39	0,000019
5	30	22,05	8,05	-0,07	0,000003
1	52405	38893,05	13512,05	-0,02	0,0004
2	1531	1293,05	238,05	0,61	0,008
3	520	444,05	76,05	0,68	0,003
4	93	80,05	13,05	0,73	0,0007
9	1	1,05	0,05	1,96	0,0004
7	4	4,05	0,05	3,31	0,0003

Tabela 3.18: Análise da variável language_count

Para os indivíduos que falam 10 idiomas, temos 1 observação, representando 0,0001% do total de observações. A proporção de bons pagadores é de 0%, com um número de bons pagadores de 0,05 e maus pagadores de 1,05. O Weight of Evidence é -4,12, indicando uma menor propensão ao bom crédito. O valor de IV é 0,0003, sugerindo uma variável de baixa importância preditiva.

Os indivíduos que falam 1 idioma (nativa) têm 52405 observações, representando 95,99

Os indivíduos bilíngues (2 idiomas) têm 1,531 observações, representando 2,80% do total, com uma proporção de bons pagadores de 84,45%. O WoE é 0,61, e o IV é 0,008, indicando uma variável de baixa importância preditiva.

Para os indivíduos trilíngues (3 idiomas), temos 520 observações, representando 0,9525% do total. A proporção de bons pagadores é de 85,38%, com um WoE de 0,68, indicando uma maior propensão ao bom crédito. O IV é 0,003, sugerindo uma importância preditiva muito baixa.

Os indivíduos poliglotas (4 ou mais idiomas) apresentam uma variação nos dados, com os que falam 4 idiomas tendo 93 observações e os que falam 9 idiomas tendo apenas 1 observação. A proporção de bons pagadores para poliglotas varia de 86,02% a 100%, com

WoE variando de 0,73 a 3,31, indicando uma alta propensão ao bom crédito. Os valores de IV também variam, mas permanecem em níveis baixos devido ao número reduzido de observações.

A análise do WoE e do IV indica que a quantidade de idiomas falados por um indivíduo é um fator significativo para a probabilidade de bom crédito, com indivíduos que falam mais idiomas (poliglotas) apresentando maior propensão ao bom crédito em comparação aos que falam menos idiomas.

Análise da variável email

A variável email indica se o indivíduo forneceu um email no perfil.

email	**n_obs**	**n_bom**	**n_ruim**	**WoE**	**IV**
0	53970	40264,05	13706,05	-0,0012	0,0002
1	622	477,05	145,05	0,11	0,0001

Tabela 3.19: Análise da variável email

A tabela 3.19 apresenta que para os indivíduos que não forneceram um email , temos 53,970 observações, representando 98,86% do total de observações. A proporção de bons pagadores é de 74,60%, com um número de bons pagadores de 40264,05 e maus pagadores (n_ruim) de 13706,05. O WoE é -0,001, indicando uma propensão ao bom crédito quase neutra. O valor de IV é 0,0002, sugerindo uma variável de importância preditiva insignificante.

Para os indivíduos que forneceram um email, temos 622 observações, representando 1,14% do total. A proporção de bons pagadores é de 76,69%, com um WoE de 0,11, indicando uma leve maior propensão ao bom crédito. O IV é 0,0001, sugerindo uma importância preditiva muito baixa.

A análise do WoE e do IV indica que fornecer ou não um email é um fator pouco significativo para a probabilidade de bom crédito, com uma leve vantagem para os indivíduos que forneceram um email. No entanto, a importância preditiva geral da variável é muito baixa.

Análise da variável phone

phone	n_obs	n_bom	n_ruim	WoE	IV
0	53555	39905,05	13650,05	-0,006	0,00003
1	1037	836,05	201,05	0,34	0,002

Tabela 3.20: Análise da variável phone

A variável phone indica se o indivíduo forneceu um telefone. A tabela 3.20 mostra várias métricas relacionadas à distribuição e qualidade de crédito dos indivíduos que forneceram ou não forneceram um telefone.

Para os indivíduos que não forneceram um telefone, temos 53555 observações, representando 98,10% do total de observações. A proporção de bons pagadores é de 74,51%, com um número de bons pagadores de 39905,05 e maus pagadores de 13650,05. O WoE é -0,006, indicando uma propensão ao bom crédito quase neutra. O valor de IV é 0,00003, sugerindo uma variável de importância preditiva insignificante.

Para os indivíduos que forneceram um telefone, temos 1037 observações, representando 1,90% do total. A proporção de bons pagadores é de 80,62%, com um WoE de 0,34, indicando uma maior propensão ao bom crédito. O IV é 0,002, sugerindo uma importância preditiva muito baixa.

A análise do WoE e do IV indica que fornecer ou não um telefone é um fator pouco significativo para a probabilidade de bom crédito, com uma vantagem para os indivíduos que forneceram um telefone. No entanto, a importância preditiva geral da variável é muito baixa.

Análise da variável contact_info

contact_info	n_obs	n_bom	n_ruim	WoE	IV
0	53089	39546,05	13543,05	-0,007	0,00005
1	1503	1195,05	308,05	0,27	0,001

Tabela 3.21: Análise da variável contact_info

A variável *contact_info* foi criada a partir das variáveis *phone* e *email*, onde um indivíduo é considerado como tendo informações de contato ($contact\_info = 1$) se forneceu

telefone ou email. A tabela 3.21 mostra várias métricas relacionadas à distribuição e qualidade de crédito dos indivíduos que forneceram ou não informações de contato.

Para os indivíduos que não forneceram informações de contato, temos 53089 observações, representando 97,25% do total de observações. A proporção de bons pagadores é de 74,49%, com um número de bons pagadores de 39546,05 e maus pagadores de 13543,05. O WoE é -0,0072, indicando uma propensão ao bom crédito quase neutra. O valor de IV é 0,00005, sugerindo uma variável de importância preditiva insignificante.

Para os indivíduos que forneceram informações de contato, temos 1503 observações, representando 2,75% do total. A proporção de bons pagadores é de 79,51%, com um WoE de 0,27, indicando uma maior propensão ao bom crédito. O IV é 0,001, sugerindo uma importância preditiva muito baixa.

A análise do WoE e do IV indica que fornecer ou não informações de contato é um fator pouco significativo para a probabilidade de bom crédito, com uma leve vantagem para os indivíduos que forneceram informações de contato. No entanto, a importância preditiva geral da variável é muito baixa.

Análise da variável nickname

nickname	**n_obs**	**n_bom**	**n_ruim**	**WoE**	**IV**
0	43258	31893,05	11365,05	-0,04	0,001
1	11334	8848,05	2486,05	0,19	0,007

Tabela 3.22: Análise da variável nickname

A variável *nickname* indica se o indivíduo forneceu um apelido. A tabela 3.22 mostra várias métricas relacionadas à distribuição e qualidade de crédito dos indivíduos que forneceram ou não forneceram um apelido.

Para os indivíduos que não forneceram um apelido, temos 43258 observações, representando 79,24% do total de observações. A proporção de bons pagadores é de 73,73%, com um número de bons pagadores de 31893,05 e maus pagadores de 11365,05. O WoE é -0,04, indicando uma leve menor propensão ao bom crédito. O valor de IV é 0,001, sugerindo uma variável de importância preditiva muito baixa.

Para os indivíduos que forneceram um apelido, temos 11334 observações, representando 20,76% do total. A proporção de bons pagadores é de 78,07%, com um WoE de 0,19, indicando uma maior propensão ao bom crédito. O IV é 0,007, sugerindo uma importância preditiva muito baixa.

A análise do WoE e do IV indica que fornecer ou não um apelido é um fator pouco significativo para a probabilidade de bom crédito, com uma leve vantagem para os indivíduos que forneceram um apelido. No entanto, a importância preditiva geral da variável é muito baixa.

Análise da variável foreigner

foreigner	**n_obs**	**n_bom**	**n_ruim**	**WoE**	**IV**
1	9204	6223,05	2981,05	-0,34	0,02
0	45388	34518,05	10870,05	0,07	0,004

Tabela 3.23: Análise da variável foreigner

A variável foreigner indica se o indivíduo é estrangeiro. A tabela 3.23 mostra que para os indivíduos estrangeiros, temos 9204 observações, representando 16,86% do total de observações. A proporção de bons pagadores é de 67,61%, com um número de bons pagadores de 6223,05 e maus pagadores de 2981,05. O WoE é -0,34, indicando uma menor propensão ao bom crédito. O valor de IV é 0,02, sugerindo uma variável de baixa importância preditiva.

Para os indivíduos não estrangeiros, temos 45388 observações, representando 83,14% do total. A proporção de bons pagadores é de 76,05%, com um WoE de 0,07, indicando uma maior propensão ao bom crédito. O IV é 0,004, sugerindo uma importância preditiva muito baixa.

A análise do WoE e do IV indica que ser estrangeiro ou não é um fator pouco significativo para a probabilidade de bom crédito, com uma vantagem para os indivíduos não estrangeiros. No entanto, a importância preditiva geral da variável é baixa.

Análise da variável profile_id

A cada indivíduo no facebook é atribuído um profile id único. A variável será utilizada

como índice para o conjunto de dados.

Análise da variável name

A variável name representa o nome completo do indivíduo, tendo sido removido do conjunto de dados para anominação dos dados.

Transformação da Variável username

A variável name representa o nome de usuário do indivíduo, tendo sido removido do conjunto de dados para anominação dos dados.

Transformação da variável hometown

as variáveis relacionadas a origem do indivíduo hometown, hometown_state w hometown_state_code foram removidas do conjunto de dados para evitar discriminações relacionadas a origem dos indivíduos, apenas a região de origem foi considera por representar a migração interna no país.

Variáveis religion e political_statement

Essas variáveis podem criar um viés potencialmente discriminatório referente às crenças e pensamento político do indivíduo no modelo.

Análise da variável education

A variável education armazena dados sobre a formação acadêmica do indivíduo. A variável foi processada e transformada em variável categórica que armazena o nível de graduação mais alta indicado. As categorias classificadas foram: doutorado, mestrado, especialização, graduado, ensino médio, ensino fundamental

education	**n_obs**	**n_bom**	**n_ruim**	**WoE**	**IV**
fundamental	24067	17174,05	6893,05	-0,16	0,01
graduado	4557	3423,05	1134,05	0,025	0,00005
medio	24870	19223,05	5647,05	0,14	0,009
doutorado	46	37,05	9,05	0,33	0,0008
mestrado	166	139,05	27,05	0,55	0,0008
especialista	886	745,05	141,05	0,58	0,004

Tabela 3.24: Análise da variável education

A variável education representa o nível de educação dos indivíduos. A tabela 3.24 mostra várias métricas relacionadas à distribuição e qualidade de crédito dos indivíduos com diferentes níveis de educação.

Para os indivíduos com educação fundamental, temos 24067 observações, representando 44,09% do total de observações. A proporção de bons pagadores é de 71,36%, com um número de bons pagadores de 17174,05 e maus pagadores de 6893,05. O WoE é -0,16, indicando uma menor propensão ao bom crédito. O valor de IV é 0,01, sugerindo uma variável de baixa importância preditiva.

Para os indivíduos graduados, temos 4557 observações, representando 8,35% do total. A proporção de bons pagadores é de 75,12%, com um WoE de 0,02, indicando uma leve maior propensão ao bom crédito. O IV é 0,0005, sugerindo uma importância preditiva insignificante.

Para os indivíduos com ensino médio, temos 24870 observações, representando 45,56% do total. A proporção de bons pagadores é de 77,29%, com um WoE de 0,14, indicando uma maior propensão ao bom crédito. IV é 0,009, sugerindo uma importância preditiva baixa.

Para os indivíduos com doutorado, temos 46 observações, representando 0,08% do total. A proporção de bons pagadores é de 80,43%, com um WoE de 0,33, indicando uma maior propensão ao bom crédito. O IV é 0,0008, sugerindo uma importância preditiva insignificante.

Para os indivíduos com mestrado, temos 166 observações, representando 0,30% do total. A proporção de bons pagadores é de 83,73%, com um WoE de 0,55, indicando uma maior propensão ao bom crédito. O IV é 0,0008, sugerindo uma importância preditiva baixa.

Para os indivíduos especialistas, temos 886 observações, representando 1,62% do total. A proporção de bons pagadores é de 84,09%, com um WoE de 0,58, indicando uma maior propensão ao bom crédito. O IV é 0,004, sugerindo uma importância preditiva baixa.

Análise da variável follower_count

follower_count	n_obs	n_bom	n_ruim	WoE	IV
0	54583	40734,05	13849,05	-0,0002	6,27
1	9	7,05	2,05	0,15	3,91

Tabela 3.25: Análise da variável follower_count

A variável follower_count representa a quantidade de seguidores que um indivíduo possui. A tabela 3.25 mostra várias métricas relacionadas à distribuição e qualidade de crédito dos indivíduos com diferentes quantidades de seguidores.

Para os indivíduos com 0 seguidores, temos 54583 observações, representando 99,98% do total de observações. A proporção de bons pagadores é de 74,63%, com um número de bons pagadores de 40734,05 e maus pagadores de 13849,05. O WoE é -0,0002, indicando uma propensão ao bom crédito quase neutra. O valor de IV é 6,27, sugerindo uma variável de importância preditiva insignificante.

Para os indivíduos com 1 seguidor, temos 9 observações, representando 0,000165% do total. A proporção de bons pagadores é de 77,78%, com um WoE de 0,15, indicando uma maior propensão ao bom crédito. O IV é 3,91, sugerindo uma importância preditiva insignificante.

Análise da variável friends_link

A variável friends_link representa a quantidade de relacionamentos abertos no facebook. a variável foi utilizada para a extração dos relacionamentos do indivíduo.

friends_link	n_obs	n_bom	n_ruim	WoE	IV
1	72	48,05	24,05	-0,38	0,0002
0	54166	40410,05	13756,05	-0,0012	0,0002
2	354	283,05	71,05	0,30	0,0005

Tabela 3.26: Análise da variável friends_link

A variável friends_link representa a quantidade de links de amigos que um indivíduo possui. A tabela 3.26 mostra várias métricas relacionadas à distribuição e qualidade de crédito dos indivíduos com diferentes quantidades de links de amigos.

Para os indivíduos da categoria 1, entre 2 e 500 links de amigos, temos 72 observações, representando 0,13% do total de observações. A proporção de bons pagadores é de 66,67%, com um número de bons pagadores de 48,05 e maus pagadores de 24,05. O WoE é -0,38, indicando uma menor propensão ao bom crédito. O valor de IV é 0,0002, sugerindo uma variável de importância preditiva muito baixa.

Para os indivíduos sem links de amigos, temos 54166 observações, representando 99,22% do total. A proporção de bons pagadores é de 74,60%, com um WoE de -0,0012, indicando uma propensão ao bom crédito quase neutra. O IV é 0,0002, sugerindo uma importância preditiva insignificante.

Para os indivíduos com 2mais de 435 links de amigos, temos 354 observações, representando 0,65% do total. A proporção de bons pagadores é de 79,94%, com um WoE de 0,30, indicando uma maior propensão ao bom crédito. O IV é 0,0005, sugerindo uma importância preditiva muito baixa.

Análise da variável facebook

A variável facebook representa o link para o profile do indivíduo. Todos os indivíduos no facebook tem um link de acesso. Com isso a variável foi considerada irrelevante e removida do modelo.

3.4.6 Relacionamentos

A estrutura de grafos em uma rede social, como o Facebook, representa a maneira como os nodos (usuários) e os relacionamentos (amizades e interações) estão conectados entre si. Para este estudo foram coletados dados do perfil de redes sociais, começando por um indivíduo, foi extraída a lista de amigos; em seguida, para cada um desses amigos, foi extraída novamente a lista de amizades. Assim, cada nó extraído possui, no máximo, três conexões diretas (dois saltos) ao individuo inicial. Cada nó representa um perfil na rede social, enquanto cada aresta representa uma relação de amizade bidirecional entre dois perfis. No conjunto de dados inicial, foram considerados apenas os dados relacionados aos indivíduos incluídos após o processamento do conjunto de dados. O grafo foi criado utilizando o *profile_id* como identificador dos nós e os relacionamentos de amizades como arestas.

Após o ajuste do conjunto de dados, foram calculadas várias medidas de centralidade, tais como *betweenness centrality*, *degree centrality*, *eigenvector centrality*, *closeness centrality*, *pagerank* e *detecção de comunidades*. Cada uma dessas variáveis oferece percepções sobre a importância e o valor de cada dado na rede.

PageRank

O **PageRank** é uma métrica que avalia a importância de um nó no grafo com base nas conexões que ele recebe. No contexto da análise de crédito, o PageRank pode identificar perfis altamente referenciados por outros perfis, indicando um nível de confiança ou influência social. Um perfil com um PageRank elevado pode ser visto como mais confiável, já que muitas pessoas importantes estão conectadas a ele, o que pode ser um fator positivo na avaliação de crédito.

A centralidade PageRank mede a importância de um nó com base na importância de seus vizinhos de entrada, com um fator de amortecimento para considerar saltos aleatórios.

Para um nó v:

$$PR(v) = \frac{1-d}{N} + d \sum_{u \in \text{vizinhos de entrada}(v)} \frac{PR(u)}{\text{grau de saída}(u)}$$

- $PR(v)$: PageRank do nó v.
- d: fator de amortecimento (tipicamente definido como 0,85), representando a probabilidade de que um nó seguirá um link de saída.
- N: número total de nós na rede.
- $\sum_{u \in \text{vizinhos de entrada}(v)}$: somatório sobre todos os nós u que têm uma aresta apontando para o nó v (vizinhos de entrada).
- $PR(u)$: PageRank do nó u.
- grau de saída(u): número de arestas de saída do nó u.

Value	All	Bad	Good	Bad Rate	D_good	D_bad	WoE	IV
series 0	27143	21413	5730	0,78	0,41	0,52	-0,23	0,02
series 1	27449	19328	8121	0,70	0,58	0,47	0,21	0,02

Tabela 3.27: Análise dos dados de Pagerank

A análise dos dados de PageRank fornecidos na tabela 3.27 revela informações importantes sobre a qualidade da segmentação dos dados em relação à taxa de eventos ruins e bons. Os dados indicam que o primeiro intervalo apresenta uma taxa de eventos ruims de 78,89%, significativamente maior em comparação com a taxa de 70,41% encontrada no segundo intervalo. Isso sugere que o primeiro intervalo de PageRank está mais associado a eventos ruins, enquanto o segundo intervalo tem uma menor associação com eventos ruins. No primeiro intervalo de PageRank, a proporção de eventos ruins é de 52,56%, enquanto a proporção de eventos bons é de 41,37%. Esse intervalo apresenta uma maior concentração de eventos ruins. Por outro lado, o segundo intervalo de PageRank tem uma proporção maior de eventos bons 58,63% em comparação com eventos ruins 47,44%, indicando uma associação mais forte com eventos bons. Ambos os intervalos de PageRank apresentam valores de IV relativamente baixos 0,02 e 0,02, sugerindo que, embora a variável PageRank tenha alguma capacidade de discriminar entre eventos ruins e bons, essa capacidade não é muito forte. Valores de IV baixos indicam que a variável, por si só, é um preditor fraco.

Centralidade de Intermediação

A centralidade de intermediação (betweenness centrality) mede quantas vezes um nó aparece nos caminhos mais curtos entre outros nós. Para a análise de crédito, perfis com alta centralidade de intermediação são críticos, ao atuarem como pontes entre diferentes grupos. Isso pode indicar que o perfil tem uma posição central em várias redes sociais, facilitando a disseminação de informações. Um indivíduo bem conectado em diferentes grupos pode ser visto como tendo uma rede de suporte diversificada e robusta, o que pode influenciar positivamente sua avaliação de crédito.

Para um nó v:

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

- $C_B(v)$: centralidade de intermediação do nó v.
- σ_{st}: número total de caminhos mais curtos do nó s para o nó t.
- $\sigma_{st}(v)$: número de caminhos mais curtos do nó s para o nó t que passam pelo nó v.
- $\sum_{s \neq v \neq t}$: somatório sobre todos os pares de nós s e t, excluindo v.

A análise dos dados de betweenness_centrality revela uma clara divisão entre intervalos associados a eventos ruins e bons. Os intervalos baixos estão fortemente associados a eventos ruins, enquanto os intervalos altos estão associados a eventos bons. Esta análise sugere que betweenness_centrality pode ser uma variável útil para discriminar entre diferentes tipos de eventos, especialmente quando combinada com outras variáveis em um modelo preditivo. Apesar disso o IV indique que cada categoria tem valor muito baixo sendo assim os dados são classificados em 7 intervalos onde o valor da informação tem relevância mínima para o modelo.

Centralidade de Proximidade

A centralidade de proximidade (closeness centrality) avalia a média da menor distância de um nó para todos os outros nós do grafo. Perfis com alta centralidade de proximidade podem alcançar rapidamente todos os outros perfis na rede. Na análise de crédito, isso pode ser interpretado como a capacidade do indivíduo de influenciar rapidamente sua rede social, o que pode ser um indicador de capacidade de mobilização e influência positiva, fatores que podem ser considerados na avaliação de risco.

A centralidade de proximidade mede o quão próximo um nó está de todos os outros nós na rede.

Para um nó v:

$$C_C(v) = \frac{1}{\sum_{u \in V} d(v,u)}$$

- $C_C(v)$: centralidade de proximidade do nó v.
- V: conjunto de todos os nós na rede.
- $d(v,u)$: distância do caminho mais curto entre os nós v e u.
- $\sum_{u \in V} d(v,u)$: soma das distâncias dos caminhos mais curtos do nó v para todos os outros nós u na rede.

Value	All	ruim	bom	WoE	IV
1,13 - 4,01	11077	8267	2810	-0,0002	8,92
5,77 - 12,04	10488	8995	1493	-0,7	8,10
-0,00 - 0,11	14534	8855	5679	0,63	1,22
0,11 - 1,13	7405	5502	1903	0,01	4,02
4,01 - 5,77	11088	9122	1966	-0,45	3,73

Tabela 3.28: Análise dos dados de Closeness Centrality

Conforme os dados da tabela 3.28 permite analisar que a taxa de eventos ruins varia significativamente entre os diferentes intervalos de closeness_centrality. Por exemplo, o intervalo de 5,77 - 12,04 tem uma taxa de eventos ruins de 85,76%, enquanto o intervalo de -0,00 - 0,11 tem uma taxa de eventos ruins de 60,9%. Isso sugere que intervalos mais altos de closeness_centrality estão mais associados a eventos ruins. No intervalo de 5,77 - 12,04, a proporção de eventos ruins é significativamente maior (22,08%) comparada à proporção de eventos bons (10,78%). Em contraste, no intervalo de -0,00 - 0,11, a proporção de eventos bons (41,00%) é maior do que a de eventos ruins (21,73%). Isso indica que intervalos mais baixos de closeness_centrality estão mais associados a eventos bons. Intervalos com valores de WoE ruins, como 5,77 - 12,04 (-0,71), indicam uma associação com eventos ruins. Em contraste, intervalos com valores de WoE bons, como -0,00 - 0,11 (0,63), indicam uma associação com eventos bons. O intervalo 1,13 - 4,01 apresenta um WoE próximo de zero (-0,0002), sugerindo pouca ou nenhuma discriminação entre eventos bons e ruins. Valores de IV variam entre os intervalos, com o maior valor encontrado no intervalo -0,00 - 0,11 (0,12), sugerindo que esse intervalo possui uma capacidade preditiva significativa. Em contraste, o intervalo 1,13 - 4,01 tem um valor de IV muito baixo (8,92), indicando pouca ou nenhuma capacidade discriminativa.

Centralidade de Grau

A centralidade de grau (degree centrality) é uma medida simples que conta o número de conexões diretas que um nó possui. No contexto da análise de crédito, isso representa o número de amigos de um perfil. Perfis com alta centralidade de grau têm muitos amigos diretos e podem ser considerados populares ou bem conectados, sugerindo um bom capital social, o que pode ser um indicador positivo na avaliação de crédito, dado que redes de

apoio social podem ajudar em tempos de dificuldade financeira.

A centralidade de grau conta o número de conexões (arestas) que um nó tem.

Para um nó v:

$$C_D(v) = \text{grau}(v)$$

- $C_D(v)$: centralidade de grau do nó v.
- grau(v): grau do nó v (número de arestas conectadas ao v).

Value	**n_obs**	**ruim**	**bom**	**WoE**	**IV**
0,06 - 0,12	48827	35851	12976	0,06	0,003
0,12 - 226,59	5765	4890	875	-0,64	0,03

Tabela 3.29: Análise dos dados de Degree Centrality

Os dados em 3.29 indicam que a taxa de eventos ruins varia significativamente entre os diferentes intervalos de degree_centrality. Por exemplo, o intervalo de 0,12 - 226,59 tem uma taxa de eventos ruins de 84,82%, enquanto o intervalo de 0,06 - 0,12 tem uma taxa de eventos ruins de 73,42%. Isso sugere que intervalos mais altos de degree_centrality estão mais associados a eventos ruins.

Distribuição de Eventos bons e ruins No intervalo de 0,12 - 226,59, a proporção de eventos ruins (ruim) é significativamente maior (12,00%) comparada à proporção de eventos bons (6,32%). Em contraste, no intervalo de 0,06 - 0,12, a proporção de eventos bons (93,68%) é muito maior do que a de eventos ruins (87,99%). Isso indica que intervalos mais baixos de degree_centrality estão mais associados a eventos bons. Intervalos com valores de WoE ruins, como 0,12 - 226,59 (-0,64), indicam uma associação com eventos ruins. Em contraste, o intervalo de 0,06 - 0,12 tem um WoE positivo (0,06), indicando uma associação com eventos bons. O valor de IV mais alto é encontrado no intervalo de 0,12 - 226,59 (0,03), sugerindo que este intervalo possui uma capacidade preditiva significativa. Em contraste, o intervalo de 0,06 - 0,12 tem um valor de IV muito baixo (0,003), indicando pouca capacidade discriminativa.

Centralidade de Autovetor

A centralidade de autovetor (eigenvector centrality) não apenas contabiliza as conexões de um nó, mas também pondera essas conexões pela importância dos nós conectados. Em um grafo de Facebook, perfis com alta centralidade de autovetor não apenas têm muitos amigos, mas têm amigos que também são bem conectados. Para a análise de crédito, isso reflete uma influência profunda e integrada na rede social, indicando que o indivíduo está conectado a outros influentes e confiáveis, o que pode ser um sinal de estabilidade e confiança, influenciando positivamente sua avaliação de crédito.

A centralidade de autovetor mede a influência de um nó na rede com base na influência de seus vizinhos.

Para um nó v:

$$C_E(v) = \frac{1}{\lambda} \sum_{u \in V} A_{uv} C_E(u)$$

- $C_E(v)$: centralidade de autovetor do nó v.
- λ: maior valor próprio da matriz de adjacência A.
- A: Matriz de adjacência do grafo, onde A_{uv} é 1 se houver uma aresta entre os nós u e v, e 0 caso contrário.
- $\sum_{u \in V}$: somatório sobre todos os nós u na rede.
- $C_E(u)$: centralidade de autovetor do nó u.

Value	**n_obs**	**n_ruim**	**n_bom**	**WoE**	**IV**
0,00 - 0,00	18248	14175	4073	-0,16	0,009
0,06 - 1032,88	10917	9348	1569	-0,70	0,08
-0,00 - 0,00	14549	8865	5684	0,63	0,12
0,00 - 0,06	10878	8353	2525	-0,11	0,002

Tabela 3.30: Análise dos dados de Eigenvector Centrality

Os dados indicam que a taxa de eventos ruins varia significativamente entre os diferentes intervalos de eigenvector_centrality. o intervalo de 0,06 - 1032,88 tem uma taxa de eventos ruins de 85,63%, enquanto o intervalo de -0,00 - 0,00 tem uma taxa de eventos ruins de 60,93%. Isso sugere que intervalos mais altos de eigenvector_centrality estão

mais associados a eventos ruins. No intervalo de 0,06 - 1032,88, a proporção de eventos ruins (ruim) é significativamente maior (22,94%) comparada à proporção de eventos bons (11,33%). Em contraste, no intervalo de -0,00 - 0,00, a proporção de eventos bons (41,04%) é maior do que a de eventos ruins (21,76%). Isso indica que intervalos mais baixos de eigenvector_centrality estão mais associados a eventos bons. Intervalos com valores de WoE ruins, como 0,06 - 1032,88 (-0,70), indicam uma associação com eventos ruins. Em contraste, o intervalo de -0,00 - 0,00 tem um WoE positivo (0,63), indicando uma associação com eventos bons. O intervalo 0,00 - 0,00 apresenta um WoE negativo menor (-0,16), sugerindo uma associação moderada com eventos ruins. O valor de IV mais alto é encontrado no intervalo de -0,00 - 0,00 (0,12), sugerindo que este intervalo possui uma capacidade preditiva significativa. Em contraste, o intervalo de 0,00 - 0,06 tem um valor de IV muito baixo (0,002), indicando pouca capacidade discriminativa.

Algoritmo de Propagação de Rótulos (Label Propagation)

O algoritmo de propagação de rótulos (Label Propagation) é um método simples e eficiente para a detecção de comunidades em redes complexas. Ele opera iterativamente, onde cada nó na rede adota o rótulo que é mais comum entre seus vizinhos imediatos. Inicialmente, cada nó é atribuído a uma comunidade distinta. Durante as iterações, os rótulos são atualizados conforme a maioria dos rótulos de seus vizinhos, promovendo a formação de comunidades coesas. Esse processo continua até que uma configuração estável seja alcançada, onde os rótulos não mudam mais significativamente. A fórmula básica para atualizar o rótulo l_i de um nó i é:

$$l_i = \arg\max_{l} \sum_{j \in \mathcal{N}(i)} \delta(l, l_j)$$

onde $\mathcal{N}(i)$ é o conjunto de vizinhos do nó i e $\delta(l, l_j)$ é a função delta de Kronecker, sendo 1 se $l = l_j$ e 0 caso contrário. O principal atrativo deste algoritmo é sua eficiência computacional, conseguindo lidar com grandes redes em tempo linear, embora possa não garantir uma resolução ideal das comunidades devido à sua natureza heurística.

Value	n_obs	n_ruim	n_bom	WoE	IV
-0,00 - 1,00	42753	31444	11309	0,05	0,002
13,00 - 50,00	5692	4276	1416	-0,02	0,00007
50,00 - 104,00	5208	4218	990	-0,37	0,01
1,00 - 13,00	939	803	136	-0,69	0,006

Tabela 3.31: Análise dos dados de Label Propagation

Os dados em 3.31 indicam que a taxa de eventos ruins varia entre os diferentes intervalos de label_propagation. Por exemplo, o intervalo de 1,00 - 13,00 tem uma taxa de eventos ruins de 85,52%, enquanto o intervalo de -0,00 - 1,00 tem uma taxa de eventos ruins de 73,55%. Isso sugere que intervalos mais altos de label_propagation estão mais associados a eventos ruins.

Distribuição de Eventos bons e ruins No intervalo de 1,00 - 13,00, a proporção de eventos ruins (ruim) é maior (1,97%) comparada à proporção de eventos bons (0,98%). Em contraste, no intervalo de -0,00 - 1,00, a proporção de eventos bons (81,65%) é maior do que a de eventos ruins (77,18%). Isso indica que intervalos mais baixos de label_propagation estão mais associados a eventos bons. Intervalos com valores de WoE ruins, como 1,00 - 13,00 (-0,69), indicam uma associação com eventos ruins. Em contraste, o intervalo de -0,00 - 1,00 tem um WoE positivo (0,05), indicando uma associação com eventos bons.O valor de IV mais alto é encontrado no intervalo de 50,00 - 104,00 (0,011), sugerindo que este intervalo possui uma capacidade preditiva significativa. Em contraste, o intervalo de 13,00 - 50,00 tem um valor de IV muito baixo (0,0007), indicando pouca capacidade discriminativa.

Algoritmo de Louvain

O algoritmo de Louvain é um método popular para a detecção de comunidades em grafo baseado na otimização da modularidade. Ele trabalha em duas fases principais: na primeira fase, cada nó é inicialmente considerado uma comunidade separada. Em seguida, cada nó é movido para a comunidade de um de seus vizinhos se isso resultar em um aumento da modularidade. Essa fase é repetida iterativamente até que nenhuma melhoria adicional possa ser feita. Na segunda fase, o grafo é reconstruído, onde as

comunidades encontradas na primeira fase são colapsadas em super nós, e o processo é repetido.

A modularidade Q é definida como:

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

onde:

- A_{ij} é a matriz de adjacência (1 se há uma aresta entre i e j, 0 caso contrário),
- k_i e k_j são os graus dos nós i e j,
- m é o número total de arestas,
- $\delta(c_i, c_j)$ é a função delta de Kronecker, que é 1 se i e j estão na mesma comunidade, 0 caso contrário.

O algoritmo de Louvain é reconhecido por sua capacidade de encontrar comunidades de alta qualidade e sua eficiência, sendo amplamente utilizado em várias aplicações de análise de redes devido à sua capacidade de escalar eficientemente para grafos grandes.

Tabela 3.32: Análise dos dados de Louvain

Value	**n_obs**	**_ruim**	**n_bom**	**WoE**	**IV**
45,00 - 54,00	4606	3710	896	-0,34	0,009
-0,00 - 11,00	6036	4551	1485	-0,04	0,0001
17,00 - 22,00	8170	5971	2199	0,07	0,0009
64,00 - 71,00	5431	4160	1271	-0,10	0,001
54,00 - 59,00	6676	5354	1322	-0,31	0,01
39,00 - 45,00	5747	4738	1009	-0,46	0,02
59,00 - 64,00	3610	2744	866	-0,07	0,0003
22,00 - 25,00	1416	1175	241	-0,50	0,005
25,00 - 39,00	6475	3932	2543	0,64	0,05
11,00 - 17,00	6425	4406	2019	0,29	0,01

Os dados indicam que a taxa de eventos ruins varia entre os diferentes intervalos de Louvain. Por exemplo, o intervalo de 22,00 - 25,00 tem uma taxa de eventos ruins de 82,98%, enquanto o intervalo de 25,00 - 39,00 tem uma taxa de eventos ruins de 60,73%. Isso sugere que alguns intervalos de Louvain estão mais associados a eventos ruins, enquanto outros estão mais associados a eventos bons. No intervalo de 22,00 - 25,00, a proporção de eventos ruins (ruim) é maior (2,88%) comparada à proporção de eventos bons (1,74%). Em contraste, no intervalo de 25,00 - 39,00, a proporção de eventos bons (18,36%) é maior do que a de eventos ruins (9,65%). Isso indica que alguns intervalos específicos de Louvain estão mais associados a eventos bons.Intervalos com valores de WoE ruins, como 22,00 - 25,00 (-0,50), indicam uma associação com eventos ruins. Em contraste, o intervalo de 25,00 - 39,00 tem um WoE positivo (0,64), indicando uma associação com eventos bons. O valor de IV mais alto é encontrado no intervalo de 25,00 - 39,00 (0,05), sugerindo que este intervalo possui uma capacidade preditiva significativa. Em contraste, o intervalo de -0,00 - 11,00 tem um valor de IV muito baixo (0,0001), indicando pouca capacidade discriminativa.

3.5 Preparação dos dados.

Antes da aplicação dos modelos de Machine Learning, foi realizada a preparação do conjunto de dados para os modelos. Esta etapa garantiu a qualidade e a adequação dos dados para análise preditiva, incluindo o balanceamento das classes, a análise da variável de interesse, a criação de variáveis fictícias, e a divisão dos dados em conjuntos de treino e teste. A variável default foi analisada para identificar a proporção de clientes inadimplentes e não inadimplentes. Esta variável, sendo categórica, indicava se um cliente havia ou não inadimplido. Observou-se que os dados estavam desbalanceados, com uma maioria de clientes inadimplentes. Esse desbalanceamento pode prejudicar a desempenho dos modelos de Machine Learning, que tendem a favorecer a classe majoritária. Para lidar com o desbalanceamento, foi utilizado o método SMOTE (Synthetic Minority Oversampling Technique). O SMOTE é uma técnica de oversampling que cria novas instâncias sintéticas para a classe minoritária (inadimplentes) ao invés de simplesmente replicar instâncias existentes. Isso ajuda a aumentar a representatividade da classe minoritária sem introduzir duplicações que poderiam levar ao overfitting.

Seleção das variáveis do modelo: Identificar as variáveis relevantes, variáveis de referência e remover variáveis insignificantes estatisticamente. Esse tipo de análise ajuda a identificar quais atributos são mais valiosos para o modelo de predição e quais podem ser considerados para exclusão ou tratamento diferenciado para melhorar o desempenho do modelo.

original_feature	**IV_sum**
age_group	1,61
betweenness_centrality	0,00001
closeness_centrality	0,31
community_label_propagation	0,03
community_Louvain	0,13
contact_info	6,90
current_state	2,002
degree_centrality	0,079
education	0,04
eigenvector_centrality	0,29
following_count	0,0002
foreigner	0,078
friend_count	0,02
friends_link	0,0009
gender	0,04
hometown_state_region	1,25
id	0,0001
iscapitalcity	0,05
job_count	0,89
language_count	0,02
marital_status	0,18
nickname	0,02
pagerank	0,10
region	3,16

Tabela 3.33: Tabela de variáveis e seus valores de IV

Os valores de IV na tabela 3.33 fornecem uma indicação de quão informativas são as variáveis em relação ao objetivo da análise. A variável contact_info apresenta o maior valor de IV (6,90), indicando uma forte capacidade preditiva. Isso sugere que as informações de contato são altamente diferenciadoras na modelagem utilizada. Da mesma forma, current_state com IV de 2,002 é outra variável com valor elevado, sugerindo que o estado atual do indivíduo é altamente relevante para a análise. Além disso, a variável region também mostra alta capacidade preditiva com um IV de 3,16, possivelmente refletindo diferenças regionais significativas que afetam a análise. A age_group com IV de 1,61 indica que a faixa etária do indivíduo é um fator preditivo importante. A hometown_state_region (IV = 1,25) e job_count (IV = 0,89) também se destacam como variáveis significativamente preditivas.

Variáveis como closeness_centrality (IV = 0,31) e eigenvector_centrality (IV = 0,29), que são métricas de centralidade de rede, indicam uma capacidade preditiva moderada. Estas métricas sugerem que a posição do indivíduo na rede social pode ter algum impacto na análise. Outras variáveis com capacidade preditiva moderada incluem marital_status (IV = 0,18), que reflete o estado civil do indivíduo, e community_Louvain (IV = 0,13), uma variável relacionada à detecção de comunidades na rede social. pagerank (IV = 0,10), que mede a importância de um nó na rede, também apresenta uma capacidade preditiva moderada.

Algumas variáveis apresentam baixa capacidade preditiva, como degree_centrality (IV = 0,07), apesar de ser uma métrica de centralidade. O status de foreigner (IV = 0,07) e a variável iscapitalcity (IV = 0,05), que indica se o indivíduo vive na capital, também têm relevância limitada. Outras variáveis com baixa capacidade preditiva incluem gender (IV = 0,04), education (IV = 0,04), e community_label_propagation (IV = 0,03). O uso de apelidos (nickname, IV = 0,02), o número de idiomas falados (language_count, IV = 0,02) e a quantidade de amigos na rede social (friend_count, IV = 0,02) são igualmente preditores de baixa capacidade. Algumas variáveis apresentam capacidade preditiva muito baixa, como friends_link (IV = 0,0009), que indica os links entre amigos na rede. A contagem de seguidores (following_count, IV = 0,0002) e a métrica de centralidade betweenness_centrality (IV = 0,0001) também mostram relevância quase nula. O identificador do indivíduo (id, IV = 0,0001) não contribui para a capacidade preditiva do modelo.

Variáveis de referência.

Para a análise dos dados utilizando variáveis categóricas, a escolha das referências é crucial para a criação de variáveis fictícias eficazes. No caso específico, as referências foram selecionadas com base no menor Weight of Evidence (WoE), que é uma métrica utilizada para medir a força de predição de uma variável categórica em relação ao evento de interesse. As categorias selecionadas como referências são as seguintes: 'age_group:56,00 - 117,00', 'marital_status:divorced', 'gender:male', 'iscapitalcity:False', 'current_state: amazonas', 'hometown_state_region:norte', 'region: north', 'language_count: poliglota', 'contact_info:False', 'nickname: True', 'foreigner: True', 'education: fundamental', 'following_count:False' 'friends_link: 0', 'friend_count: 2', 'job_count:1', 'betweenness_centrality: (-0,232, 77,348]', 'closeness_centrality: -0,00 - 0,11', 'degree_centrality:0,06 - 0,12', 'eigenvector_centrality:-0,00 - 0,00', 'community_label_propagation:-0,00 - 1,00', 'community_Louvain:25,00 - 39,00', 'pagerank:-0,00 - 0,00'. A remoção de uma categoria em cada variável é necessária para evitar a multicolinearidade, que ocorre quando uma categoria pode ser perfeitamente predita pelas outras, causando problemas na estimativa dos coeficientes dos modelos de regressão. Dessa forma, ao definir uma categoria de referência, garantimos que cada variável categórica contribua de maneira independente para o modelo, proporcionando uma análise mais robusta e interpretações mais precisas.

Identificação das Instâncias da Classe Minoritária: As instâncias pertencentes à classe inadimplente foram identificadas. O conjunto de dados final possui 102 variáveis em 54591 itens.

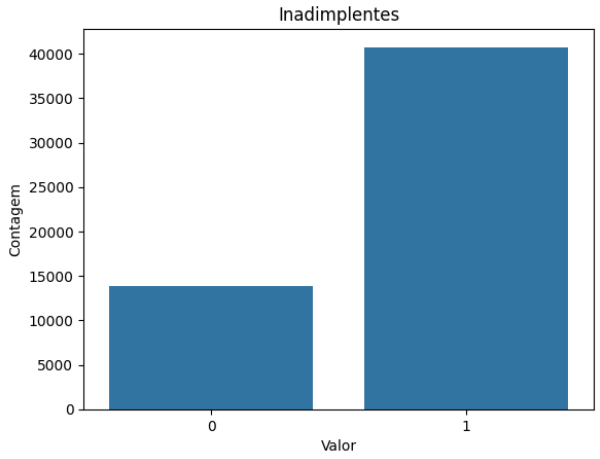

Figura 3.6: Proporção de adimplentes e Inadimplentes

A imagem acima 3.6 apresenta a diferença entre as divisões da variavel objetivo *default*. Conforme é possível entender, o conjunto de dados está desbalanceado e representa melhormente a classe referente a inadimplência.

Divisão dos Dados em Treino e Teste. Após a preparação inicial, os dados foram divididos em conjuntos de treino e teste. Essa divisão é fundamental para avaliar a desempenho dos modelos de Machine Learning de maneira justa e robusta. Os dados foram divididos de forma que 70% dos dados fossem usados para treinamento e 30% para teste. A divisão foi realizada de maneira aleatória para garantir que ambos os conjuntos sejam representativos da distribuição geral dos dados. Novas instâncias sintéticas foram geradas ao combinar instâncias existentes da classe minoritária de maneira a criar pontos de dados novos que representassem a variabilidade da classe. Foi aplicado o Método SMOTE para balanceamento das classes. As instâncias sintéticas foram combinadas com os dados originais da classe majoritária (não inadimplentes) para criar um conjunto de dados balanceado.

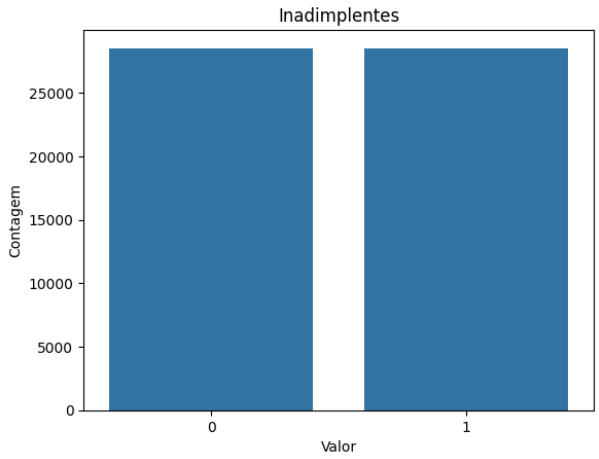

Figura 3.7: Classes da variável default balanceadas.

A figura 3.7 apresenta o balanceamento das classes após a aplicação de SMOTE. O conjunto de dados representa igualmente ambas classes da variavel objetivo.

Essa abordagem garantiu que os modelos de Machine Learning fossem treinados e avaliados em dados distintos, permitindo uma avaliação precisa de sua capacidade de generalização.

3.6 Modelagem.

3.6.1 Regressão Logística

A regressão logística é apropriada para problemas de classificação binária, onde a variável dependente é categórica. Na análise de crédito, é utilizada para prever a probabilidade de um cliente ser inadimplente (1) ou não (0).

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{3.1}$$

Onde:

- p é a probabilidade de inadimplência.
- $\beta_0, \beta_1, \beta_2, \cdots, \beta_n$ são os coeficientes das variáveis independentes.

3.6.2 Árvore de Decisão

A árvore de decisão é um modelo de aprendizado supervisionado que usa uma estrutura em árvore para tomar decisões com base nas características dos dados. É particularmente útil para interpretar e visualizar decisões de classificação. A árvore de decisão divide recursivamente dos dados em subconjuntos mais homogêneos e utiliza critérios como Gini ou Entropia para escolher as divisões.

3.6.3 Random Forest

Random Forest é um modelo de aprendizado supervisionado que utiliza um conjunto de árvores de decisão para melhorar a precisão da classificação e reduzir o overfitting. Este método é particularmente eficaz para lidar com grandes volumes de dados e variáveis altamente correlacionadas.

O modelo Random Forest funciona criando várias árvores de decisão durante o treinamento e, em seguida, combinando os resultados dessas árvores para produzir uma previsão

final. Cada árvore é construída a partir de um subconjunto aleatório dos dados e das características, ajudando a diversificar as previsões e reduzir o risco de overfitting.

3.6.4 Gradient Boosting

Gradient Boosting é uma técnica de *ensemble learning* que combina vários modelos fracos (normalmente árvores de decisão) para criar um modelo forte. Ele otimiza a função de perda iterativamente, ajustando modelos simples para corrigir os erros dos modelos anteriores. O algoritmo de gradient boosting Inicializa o modelo com uma predição constante. Após, ajusta um modelo simples aos resíduos (erros) do modelo atual. Atualiza o modelo adicionando o novo modelo ajustado e repete até atingir um número de iterações ou um nível de erro aceitável.

3.7 Validação

3.7.1 Acurácia

A acurácia é a proporção de previsões corretas entre o total de previsões realizadas.

$$\text{Acurácia} = \frac{\text{Número de previsões corretas}}{\text{Total de previsões}} \tag{3.2}$$

3.7.2 Matriz de Confusão

A matriz de confusão é uma ferramenta que permite visualizar o desempenho do modelo, mostrando a relação entre as previsões verdadeiras e as previsões falsas.

Estrutura:

- Verdadeiros bons (**VP**): número de previsões positivas corretas.
- Verdadeiros ruins (**VN**): número de previsões negativas corretas.
- Falsos bons (**FP**): número de previsões positivas incorretas.

- Falsos ruins (**FN**): número de previsões negativas incorretas.

3.7.3 Coeficiente de Gini

O coeficiente de Gini é uma medida de desigualdade que pode ser utilizada para avaliar a capacidade de discriminação de um modelo.

$$\text{Gini} = 1 - \sum_{i=1}^{n}(P_i + P_{i-1})(R_i - R_{i-1}) \tag{3.3}$$

Onde:

- P_i é a proporção acumulada dos bons até o ponto i.
- R_i é a proporção acumulada dos ruins até o ponto i.

3.7.4 Teste de Kolmogorov-Smirnov (KS)

O teste de Kolmogorov-Smirnov mede a distância máxima entre as distribuições acumuladas de duas amostras. No contexto da análise de crédito, ele compara a distribuição dos *scores* para os eventos de inadimplência e não inadimplência.

$$\text{KS} = \max |F_1(x) - F_2(x)| \tag{3.4}$$

Onde:

- $F_1(x)$ é a função de distribuição acumulada dos bons.
- $F_2(x)$ é a função de distribuição acumulada dos ruins.

3.8 Tomada de decisão.

Régua de corte para aprovação de crédito.

Os modelos de aprendizado de máquina fornecem dois resultados importantes. O primeiro, a probabilidade de uma entrada pertencer a uma determinada classe, e segundo,

a classe prevista com base em um limite definido. A função predizer probabilidades retorna as probabilidades estimadas para cada classe, enquanto a função predizer retorna a classe prevista diretamente, baseada em um limiar padrão (threshold), geralmente 0,5. Se a probabilidade prevista para a classe positiva for maior ou igual a 0,5, a função predizer retorna 1 (classe positiva); caso contrário, retorna 0 (classe negativa). Um bom escore de crédito deve refletir claramente a probabilidade do indivíduo pertencer à classe inadimplente. Normalmente no mercado esta probabilidade é atribuída para quanto menor for o escore maior a probabilidade de pertencer à classe inadimplente.

3.8.1 Curva ROC

A Curva ROC (Receiver Operating Characteristic) é uma ferramenta gráfica usada para avaliar o desempenho de modelos de classificação binária. A curva ROC é criada plotando a taxa de verdadeiros bons (**TPR**) contra a taxa de falsos bons (**FPR**) em diferentes **thresholds** de classificação.

- **Taxa de Verdadeiros bons (TPR)**, também conhecida como sensibilidade ou recall, é a razão entre verdadeiros bons e o total de bons reais:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

onde TP são os verdadeiros bons e FN são os falsos ruins.

- Taxa de Falsos bons (FPR), também conhecida como taxa de alarme falso, é a razão entre falsos bons e o total de ruins reais:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

onde FP são os falsos bons e TN são os verdadeiros ruins.

A Curva ROC mostra a trade-off entre a sensibilidade e a especificidade (1 - FPR) para todos os thresholds possíveis. O ponto ideal em uma Curva ROC é aquele mais próximo do canto superior esquerdo, indicando uma alta taxa de verdadeiros bons e uma baixa taxa de falsos bons.

A Área Sob a Curva (AUC) da ROC é uma métrica que quantifica o desempenho geral do modelo. Um valor de AUC mais próximo de 1 indica um modelo excelente,

enquanto um valor de AUC próximo de 0,5 indica um modelo que não é melhor do que uma classificação aleatória.

Calculamos a curva ROC e extraímos as taxas de falsos bons (FPR), as taxas de verdadeiros bons (TPR) e os thresholds correspondentes:

Definimos uma função que atribui um valor de 1 se uma probabilidade prevista for maior que o parâmetro p, sendo um limite, e um valor de 0, se não for. Em seguida, soma os valores de 1. Assim, se dados quaisquer valores percentuais, a função retornará o número de linhas com probabilidades estimadas maiores que o limite.

1. **Número de aprovados (*n_approved*)**:

$$n_{\text{aprovados}}(p) = \sum_{i=1}^{N} \mathbb{1}\left(\hat{y}_i \geq p\right)$$

onde $\mathbb{I}\left(\hat{y}_i \geq p\right)$ é 1 se $\hat{y}_i \geq p$ e 0 caso contrário, e N é o número total de solicitações.

Supondo que todas as solicitações de crédito acima de uma determinada probabilidade de serem 'boas' serão aprovadas, quando aplicamos a função 'n_aprovados' a um limite, ela retornará o número de solicitações aprovadas. Portanto, aqui calculamos o número de solicitações aprovadas para todos os limites.

$$\text{n\_aprovados}(p)$$

Em seguida, calculamos o número de solicitações rejeitadas para cada limite. É a diferença entre o número total de solicitações e as solicitações aprovadas para aquele limite.

2. **Número de rejeitados (*n_rejected*)**:

$$\text{n\_rejeitados}(p) = N - \text{n\_aprovados}(p)$$

A taxa de aprovação é igual à razão entre as solicitações aprovadas e todas as solicitações.

3. **Taxa de aprovação (*approval rate*)**:

$$\text{taxa\_aprovacao}(p) = \frac{\text{n\_aprovados}(p)}{N}$$

A taxa de rejeição é igual a um menos a taxa de aprovação.

4. **Taxa de rejeição (*rejection rate*)**:

$$\text{taxa\_rejeicao}(p) = 1 - \text{taxa\_aprovacao}(p)$$

3.9 Pontuação de crédito

A pontuação de crédito traduz a uma forma legível para as pessoas a probabilidade do indivíduo pertencer à classe inadimplente. Para isso, deve-se normalizar a probabilidade em uma pontuação de crédito. A primeira etapa deve ser definir uma pontuação máxima e mínima.

$$\text{max\_score} = 1000 \tag{3.5}$$

$$\text{min\_score} = 0 \tag{3.6}$$

A segunda etapa é extrair a importância das variáveis. Diferentes modelos apresentam esta importância de distintas formas. A Regressão Logística apresenta os coeficientes e o intercept através dos métodos `coef_` e `intercept_`. No caso de Gradient Boosting, Random Forest e Decision Tree, o método `feature_importances_` apresenta a importância de cada variável para o modelo.

$$\text{scores} = X \times Fi \tag{3.7}$$

Onde X é o conjunto de dados originais com as categorias de referência e Fi é a soma das variáveis de cada uma das variáveis pelo valor da importância ou coeficiente da variável, mais o intercepto (valor do coeficiente quando o valor de todas as variáveis é zero). Conforme o conjunto de dados utilizados no modelo possui variáveis binárias, a soma das importâncias é igual à soma da multiplicação do valor de cada variável pelo valor de sua importância. Logo após, o valor é normalizado para o intervalo entre o score mínimo e máximo.

Para garantir que os scores estivessem em um intervalo padronizado (0 a 1000), normalizamos os scores calculados utilizando a função:

$$\text{normalized\_scores} = 1000 \times \frac{\text{scores} - \text{min\_score}}{\text{max\_score} - \text{min\_score}} \tag{3.8}$$

3.9.1 Conversão de Scores para Probabilidades e Vice-Versa

Para transformar os scores em probabilidades de crédito e vice-versa, utilizamos uma base de score (*BaseScore*) e um fator (*Factor*) para encontrar o *log_odds*, sendo o logaritmo da razão entre a probabilidade de um evento ocorrer e a probabilidade de não ocorrer.

$$\text{BaseScore} = \frac{\text{max\_score}}{2} \tag{3.9}$$

$$\text{Factor} = \frac{20}{\log(2)} \tag{3.10}$$

$$\text{log\_odds} = \frac{\text{Score} - \text{BaseScore}}{\text{Factor}} \tag{3.11}$$

No caso da regressão logística, para encontrar o *log_odds* deve ser adicionado o valor do intercepto à soma dos coeficientes pelo valor da variável:

$$\text{Log-Odds} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \tag{3.12}$$

Então, as funções de conversão de probabilidade para score e vice-versa são:

$$\text{proba} = \frac{\exp(\text{log\_odds})}{1 + \exp(\text{log\_odds})} \tag{3.13}$$

Para encontrar o *log_odds* a partir da probabilidade:

$$\text{log\_odds} = \log\left(\frac{\text{proba}}{1 - \text{proba}}\right) \tag{3.14}$$

E, finalmente, para converter o *log_odds* para score:

$$\text{Score} = \text{log\_odds} \times \text{Factor} + \text{BaseScore} \tag{3.15}$$

Capítulo 4

RESULTADOS

Este capítulo apresenta os resultados obtidos a partir da aplicação do modelo de análise de crédito baseado em dados de redes sociais. Os resultados são discutidos em termos de desempenho preditivo, relevância das variáveis e comparação com modelos tradicionais de análise de crédito.

4.1 Análise das Variáveis e sua Contribuição para os Pilares da Avaliação de Crédito

A tabela abaixo resume as variáveis utilizadas na análise e indica em quais dos cinco pilares tradicionais da avaliação de crédito elas podem colaborar:

Variável	**Pilar**	**Descrição**
education	Capacidade	Indica o nível de educação do indivíduo, correlacionado com melhores oportunidades de emprego e capacidade de geração de renda.
job_count	Capacidade	Número de empregos que o indivíduo possui, indicando a estabilidade e a capacidade de geração de renda.
Contínua na próxima página		

Tabela 4.1 – Continuação da página anterior

Variável	**Pilar**	**Descrição**
following_count	Capital	Número de seguidores em redes sociais, utilizado como proxy para o capital social, indicando suporte financeiro e networking.
friend_count	Capital	Número de amigos em redes sociais, também utilizado como proxy para o capital social.
hometown_state_region	Condições	Região de origem do indivíduo, influenciando as condições econômicas e a estabilidade financeira.
current_state	Condições	Estado atual de residência, relevante para avaliar as condições econômicas regionais.
region	Condições	Região de residência, influenciando as oportunidades econômicas e a estabilidade financeira.
iscapitalcity	Condições	Indica se o indivíduo reside em uma capital, geralmente associada a melhores oportunidades econômicas.
community_louvain	Colateral	Indica a integração em comunidades, funcionando como colateral social.
community_label_propagation	Colateral	Similar ao community_louvain, indica a participação em comunidades.
degree_centrality	Colateral	Medida de centralidade em redes sociais, indicando a influência e o suporte social.
betweenness_centrality	Colateral	Outra medida de centralidade, indicando o papel de ponte entre diferentes grupos na rede social.

Contínua na próxima página

Tabela 4.1 – Continuação da página anterior

Variável	**Pilar**	**Descrição**
closeness_centrality	Colateral	Medida de quão próximo o indivíduo está de outros na rede social, indicando integração social.
eigenvector_centrality	Colateral	Medida de influência na rede social, considerando as conexões dos amigos do indivíduo.
marital_status	Caráter	Estado civil, correlacionado com a estabilidade pessoal e comportamento financeiro.
gender	Caráter	Gênero do indivíduo, utilizado para inferir padrões de comportamento financeiro.
age_group	Caráter	Grupo etário, relacionado à estabilidade financeira e padrões de comportamento.
foreigner	Caráter	Indica se o indivíduo é estrangeiro, afetando a avaliação de risco devido a desafios adicionais.

Tabela 4.1: Contribuição das variáveis para os pilares da avaliação de crédito.

A tabela 4.1 apresenta como os diferentes pilares foram atendidos pelas variáveis analisadas e quais áreas podem necessitar de melhorias ou dados adicionais.

4.1.1 Pilares Bem Atendidos

1. **Capacidade**: As variáveis education e job_count forneceram fortes indicadores da capacidade financeira dos indivíduos, alinhando-se bem com a avaliação tradicional de crédito. Indivíduos com maior nível educacional e mais empregos demonstraram uma capacidade de geração de renda mais robusta.

2. **Capital**: Utilizando following_count e friend_count como proxies para capital social, conseguimos inferir o suporte financeiro e o networking dos indivíduos. A análise mostrou que esses elementos de capital social estão conforme a avaliação tradicional de capital financeiro.

3. **Colateral**: As medidas de centralidade (degree_centrality, betweenness_centrality, closeness_centrality, eigenvector_centrality) e a participação em comunidades (community_louvain, community_label_propagation) indicaram um forte suporte social, funcionando como uma forma de colateral social. Este aspecto complementa a análise tradicional de ativos tangíveis.

4.1.2 Pilares Moderadamente Atendidos

1. **Condições**: As variáveis relacionadas à localização (hometown_state_region, current_state, region, iscapitalcity) forneceram percepções sobre as condições econômicas e a estabilidade financeira. No entanto, a análise poderia ser mais robusta com a inclusão de dados econômicos regionais mais detalhados, como taxas de desemprego e crescimento econômico local.

2. **Caráter**: As variáveis marital_status, gender, age_group e foreigner ajudaram a inferir aspectos do caráter dos indivíduos. Embora úteis, estas variáveis poderiam ser complementadas com dados adicionais sobre o histórico de crédito e comportamento financeiro dos indivíduos para uma avaliação mais completa do caráter.

4.2 Análise das Correlações das Variáveis com a Variável Default

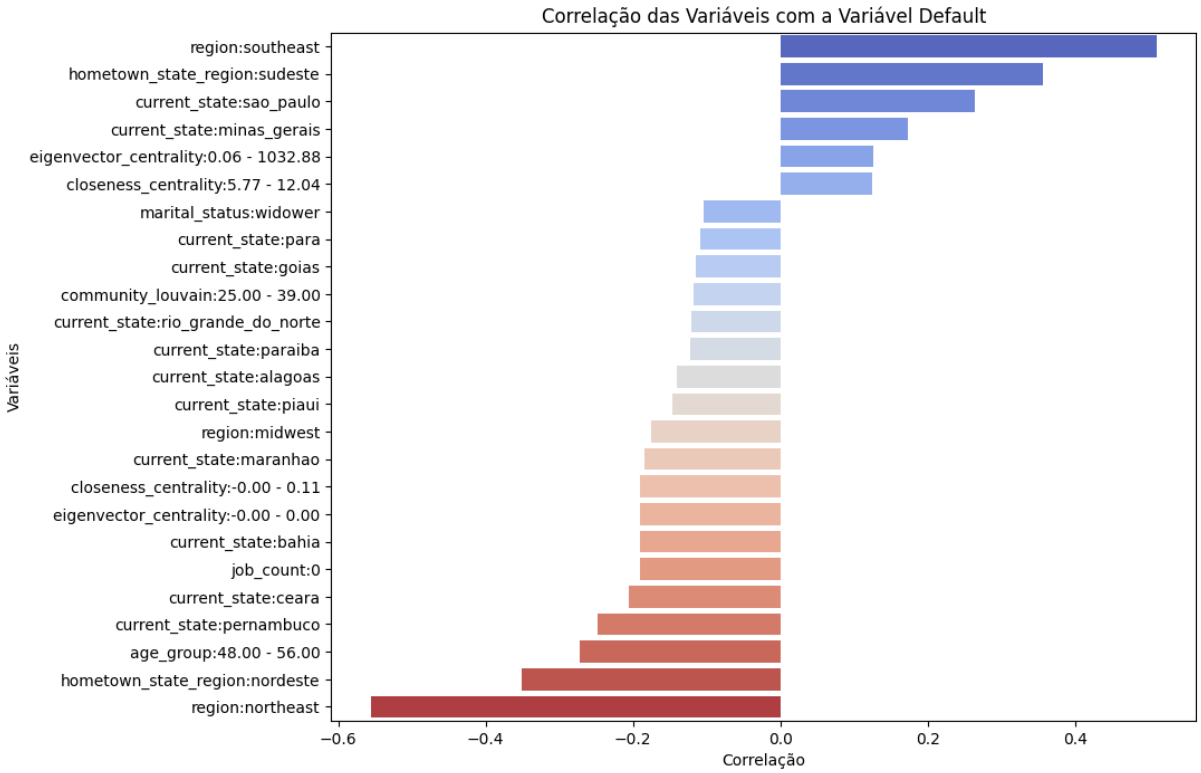

Figura 4.1: Correlação das variáveis com a variavel dependente Default

A figura 4.1 apresenta a correlação entre as variáveis independentes e a variavel de pendente default. A figura filtra apenas as variáveis com correlação relevante superior a 0,1.

4.2.1 Correlações Positivas

As variáveis com correlações positivas indicam que, conforme o valor dessas variáveis aumenta, a probabilidade de um cliente estar inadimplente (*default*) também aumenta.

- **region:southeast** (Correlação: ∼0,55)
  - **Descrição:** Indica se o cliente está localizado na região sudeste do Brasil.
  - **Interpretação:** A forte correlação positiva sugere que clientes na região sudeste têm uma maior probabilidade de inadimplência. Isso pode estar relacio-

nado a fatores econômicos regionais, como custo de vida mais alto ou políticas de crédito específicas da região.

- **hometown_state_region:sudeste** (Correlação: ∼0,50)
  - **Descrição:** Indica se o cliente nasceu na região sudeste.
  - **Interpretação:** Clientes que nasceram na região sudeste também mostram uma alta probabilidade de inadimplência, reforçando a hipótese de que fatores regionais influenciam significativamente a capacidade de pagamento.
- **current_state:sao_paulo** (Correlação: ∼0,48)
  - **Descrição:** Indica se o cliente atualmente reside no estado de São Paulo.
  - **Interpretação:** A correlação positiva moderada sugere que residentes de São Paulo têm uma maior chance de inadimplência. O alto custo de vida e a intensa competição no mercado de trabalho podem ser fatores contribuintes.
- **current_state:minas_gerais** (Correlação: ∼0,30)
  - **Descrição:** Indica se o cliente atualmente reside no estado de Minas Gerais.
  - **Interpretação:** Clientes em Minas Gerais também mostram uma correlação positiva com a inadimplência, embora menos acentuada que São Paulo.
- **eigenvector_centrality:0,06 - 1032,88** (Correlação: ∼0,28)
  - **Descrição:** Representa a centralidade de autovetor do cliente em uma rede social ou de influência.
  - **Interpretação:** Clientes com alta centralidade de autovetor são mais propensos a serem inadimplentes. Isso pode indicar que, apesar de terem um papel central em suas redes, eles podem estar sobrecarregados financeiramente.
- **closeness_centrality:5,77 - 12,04** (Correlação: ∼0,25)
  - **Descrição:** Mede a proximidade de um cliente aos outros na rede.
  - **Interpretação:** Clientes mais centralizados e acessíveis na rede têm uma maior probabilidade de inadimplência.

4.2.2 Correlações Negativas

As variáveis com correlações negativas indicam que, conforme o valor dessas variáveis aumenta, a probabilidade de um cliente estar inadimplente diminui.

- **region:northeast** (Correlação: ∼-0,55)
  - **Descrição:** Indica se o cliente está localizado na região nordeste do Brasil.
  - **Interpretação:** A forte correlação negativa sugere que clientes na região nordeste têm uma menor probabilidade de inadimplência. Isso pode refletir políticas de crédito mais conservadoras ou um menor custo de vida.
- **hometown_state_region:nordeste** (Correlação: ∼-0,45)
  - **Descrição:** Indica se o cliente nasceu na região nordeste.
  - **Interpretação:** Clientes que nasceram na região nordeste também mostram uma menor probabilidade de inadimplência, reforçando a influência de fatores regionais.
- **age_group:48,00 - 56,00** (Correlação: ∼-0,38)
  - **Descrição:** Indica se o cliente está na faixa etária de 48 a 56 anos.
  - **Interpretação:** Clientes nessa faixa etária têm uma menor probabilidade de inadimplência, possivelmente devido a uma maior estabilidade financeira e experiência na gestão de dívidas.
- **closeness_centrality:-0,00 - 0,00** (Correlação: ∼-0,30)
  - **Descrição:** Representa uma centralidade de proximidade muito baixa.
  - **Interpretação:** Clientes com valores baixos de centralidade de proximidade tendem a ser menos inadimplentes, o que pode indicar uma menor exposição a influências financeiras negativas.
- **job_count:0** (Correlação: ∼-0,30)
  - **Descrição:** Indica clientes que estão empregados.

- **Interpretação:** Clientes empregados têm uma menor probabilidade de inadimplência, destacando a importância da renda estável para a capacidade de pagamento.

- **current_state:bha** (Correlação: ∼-0,25)

  - **Descrição:** Indica se o cliente reside na Bahia.

  - **Interpretação:** Residentes da Bahia têm uma menor probabilidade de inadimplência, possivelmente devido a fatores econômicos regionais ou políticas locais de crédito.

A análise das correlações das variáveis com a variável *default* revela percepções importantes sobre os fatores que influenciam a inadimplência. As variáveis geográficas, estado civil, idade e centralidade em redes sociais têm correlações significativas com a probabilidade de inadimplência, sugerindo que esses fatores devem ser cuidadosamente considerados no desenvolvimento de modelos preditivos de crédito. Essas correlações fornecem uma base sólida para ajustar estratégias de concessão de crédito e melhorar a gestão de risco financeiro.

4.3 Desempenho preditivo

4.3.1 Modelos Utilizados

Foram utilizados quatro modelos de aprendizado de máquina para a análise de crédito: Regressão Logística, Árvore de Decisão, Floresta Randômica e Gradient Boosting. A Tabela 4.2 resume as métricas de desempenho dos quatro modelos avaliados: Regressão Logística, Gradient Boosting, Random Forest e Árvore de Decisão. As métricas consideradas foram a acurácia, F1-Score, Recall-Score e ROC AUC Score.

Modelo	Acurácia	F1-Score	Recall-Score	ROC AUC Score
Logistic Regression	0,96	0,97	0,97	0,99
Gradient Boosting	0,92	0,95	0,97	0,99
Random Forest	0,96	0,97	0,98	0,99
Decision Tree	0,97	0,98	0,97	0,96

Tabela 4.2: Métricas de desempenho dos modelos de aprendizado de máquina

Acurácia A acurácia é a proporção de previsões corretas feitas pelo modelo em relação ao total de casos. A Árvore de Decisão obteve a maior acurácia (0,97), seguida pela Regressão Logística (0,96) e Random Forest (0,96). O Gradient Boosting apresentou a menor acurácia (0,92) entre os modelos testados.

F1-Score O F1-Score é a média harmônica entre a precisão e o recall, fornecendo uma medida equilibrada das duas métricas. A Árvore de Decisão alcançou o maior F1-Score (0,98), com a Regressão Logística logo atrás (0,97). O Random Forest teve um desempenho ligeiramente inferior (0,97) e o Gradient Boosting apresentou o menor F1-Score (0,95).

Recall-Score O Recall-Score, ou sensibilidade, mede a capacidade do modelo de identificar corretamente os casos bons. O Random Forest teve o maior Recall-Score (0,98), seguido de perto pelo Gradient Boosting (0,97) e Árvore de Decisão (0,97). A Regressão Logística apresentou um Recall-Score de 0,97, sendo a menor entre os modelos.

ROC AUC Score O ROC AUC Score avalia a capacidade do modelo de distinguir entre as classes positiva e negativa. A Regressão Logística obteve o maior ROC AUC Score (0,99), demonstrando excelente desempenho em termos de separabilidade das classes. O Gradient Boosting (0,99) e Random Forest (0,99) também apresentaram altos valores de ROC AUC. A Árvore de Decisão teve o menor ROC AUC Score (0,96) entre os modelos avaliados.

4.3.2 Curva ROC e Matriz de Confusão

A seguir, apresentamos as curvas ROC dos modelos de Regressão Logística, Gradient Boosting, Random Forest e Decision Tree, que obtiveram os melhores desempenhos. As curvas ROC são fundamentais para avaliar a capacidade discriminativa dos modelos. Além disso, apresentamos as matrizes de confusão para um entendimento mais detalhado dos acertos e erros cometidos pelos modelos.

Curvas ROC

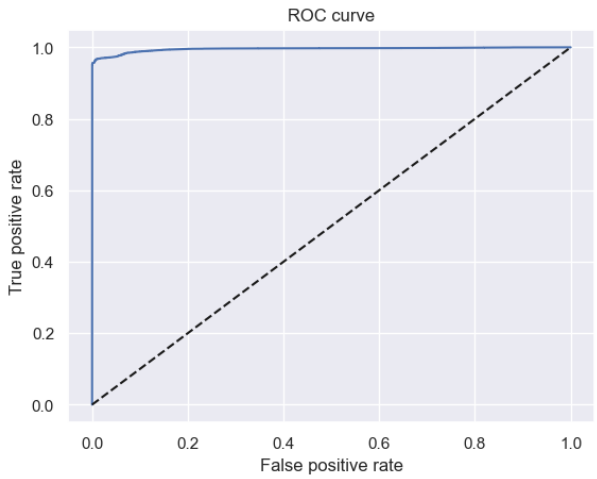

Figura 4.2: Curva ROC - Regressão Logística

A Figura 4.2 mostra a curva ROC para o modelo de Regressão Logística. A área sob a curva (AUC) é uma métrica importante para avaliar a capacidade do modelo de distinguir entre as classes positiva e negativa. A Regressão Logística apresentou uma alta AUC, indicando um bom desempenho discriminativo.

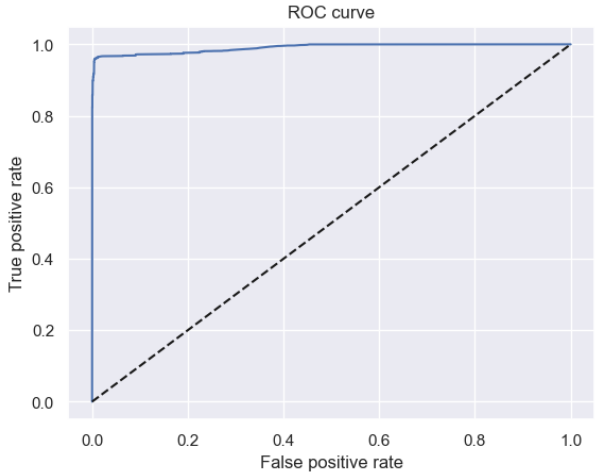

Figura 4.3: Curva ROC - Gradient Boosting

A Figura 4.3 apresenta a curva ROC para o modelo Gradient Boosting. Assim como na Regressão Logística, o Gradient Boosting também mostrou uma alta AUC, refletindo sua eficácia na separação das classes.

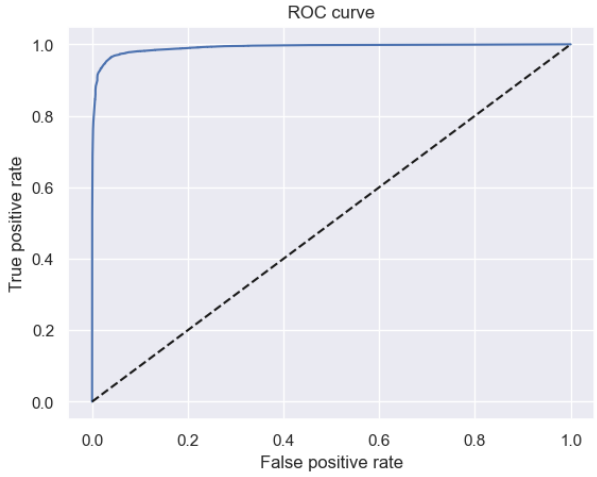

Figura 4.4: Curva ROC - Random Forest

A Figura 4.4 ilustra a curva ROC para o modelo Random Forest. Este modelo também apresentou uma alta AUC, demonstrando sua habilidade em distinguir entre as classes.

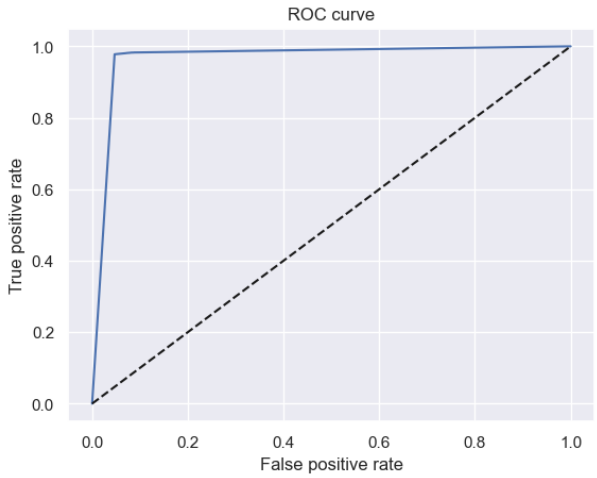

Figura 4.5: Curva ROC - Decision Tree

A Figura 4.5 exibe a curva ROC para o modelo Decision Tree. A AUC deste modelo é ligeiramente inferior às dos outros modelos apresentados, mas ainda mostra uma boa capacidade discriminativa.

Matrizes de Confusão

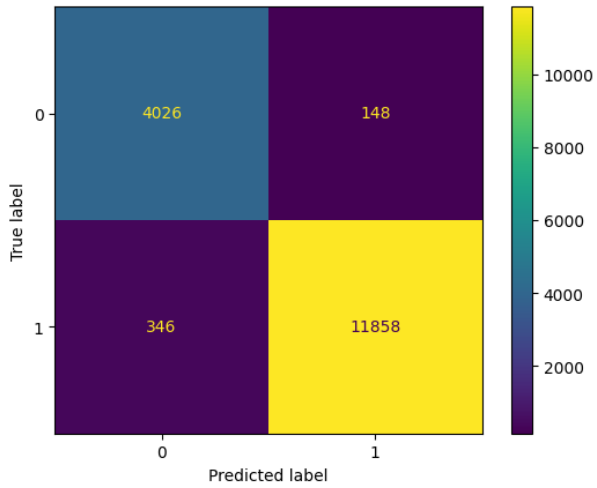

Figura 4.6: Matriz de confusão Regressão Logística

A Figura 4.6 mostra a matriz de confusão para o modelo de Regressão Logística. Nesta matriz, podemos observar o número de verdadeiros bons, verdadeiros ruins, falsos bons e falsos ruins, o que nos permite avaliar a precisão e o recall do modelo.

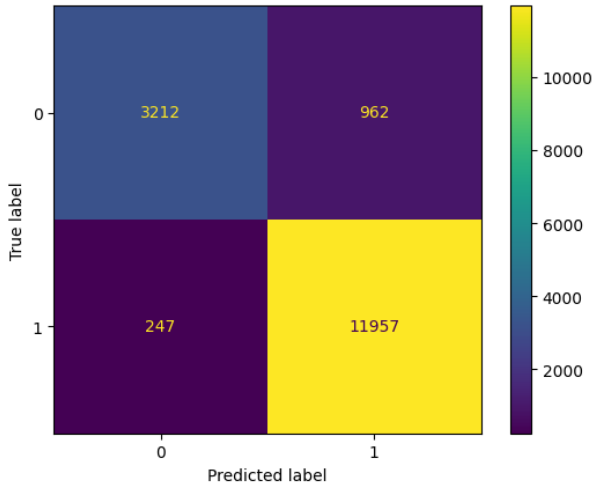

Figura 4.7: Matriz de confusão Gradient Boosting

A Figura 4.7 apresenta a matriz de confusão para o modelo Gradient Boosting. Esta matriz ajuda a entender como o modelo performa em termos de classificação correta e erros de classificação.

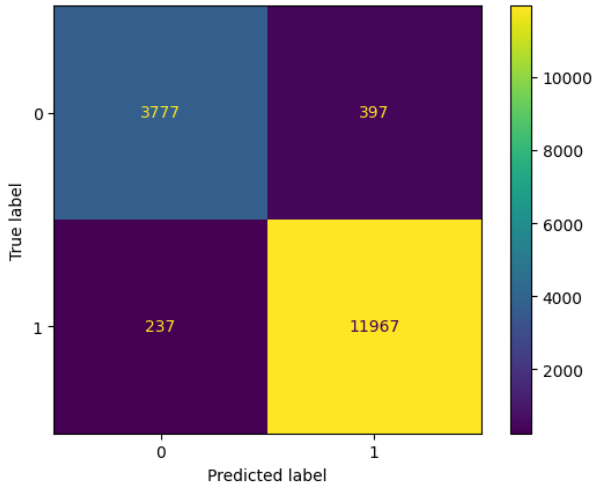

Figura 4.8: Matriz de confusão Random Forest

A Figura 4.8 exibe a matriz de confusão para o modelo Random Forest. Assim como as matrizes anteriores, esta permite avaliar o desempenho do modelo em termos de acertos e erros.

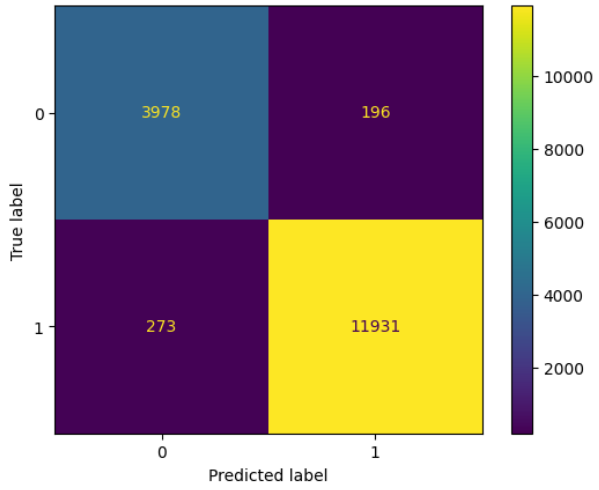

Figura 4.9: Matriz de confusão Decision Tree

A Figura 4.9 mostra a matriz de confusão para o modelo Decision Tree. A análise desta matriz fornece percepções sobre a precisão e recall deste modelo específico.

Análise dos gráficos de resultados

A seguir, apresentamos os gráficos de resultados dos modelos de aprendizado de máquina utilizados em nossa pesquisa. Os gráficos são divididos em duas categorias principais: gráficos de Gini e gráficos de Kolmogorov-Smirnov para cada modelo avaliado. Estes gráficos são importantes para avaliar o desempenho e a capacidade discriminativa dos modelos.

Regressão Logística

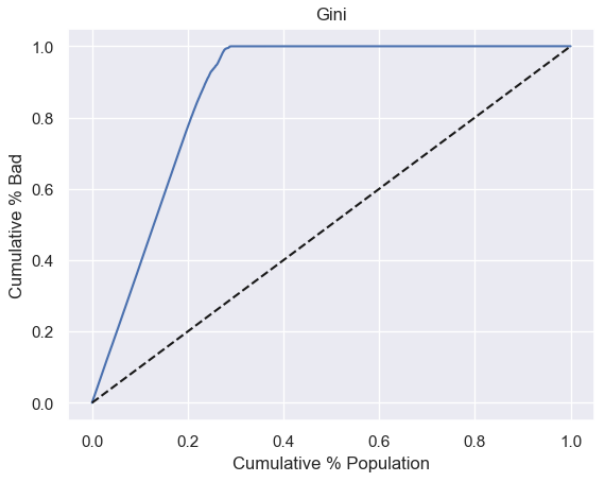

Figura 4.10: Gini Logistic Regression

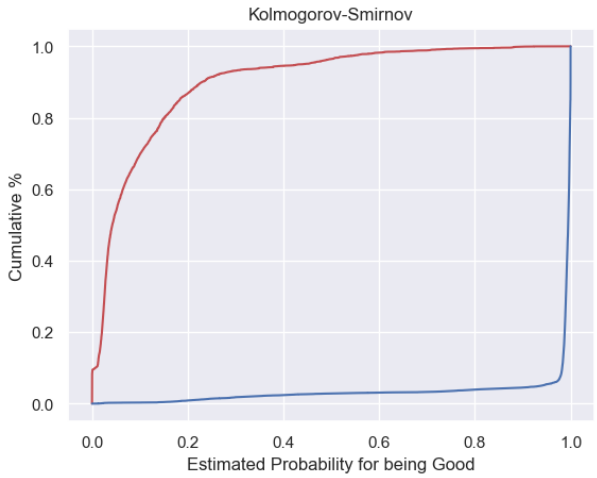

Figura 4.11: Kolmogorov Logistic Regression

Os gráficos de Gini e Kolmogorov-Smirnov para a Regressão Logística, apresentados nas Figuras 4.10 e 4.11, demonstram a capacidade do modelo em diferenciar entre as classes. A curva Gini do modelo de Regressão Logística apresenta uma excelente capacidade discriminatória. A curva se afasta significativamente da linha de igualdade, especialmente nos primeiros 20% da população. Isso indica que o modelo de Regressão Logística é altamente eficiente em identificar a maioria dos "maus"em uma pequena fração da população. A grande área entre a curva e a linha diagonal sugere um coeficiente de Gini alto, confirmando a eficácia do modelo na discriminação entre classes positivas e negativas. A curva Kolmogorov-Smirnov do modelo de Regressão Logística apresenta uma boa capacidade discriminatória. A linha vermelha (bons) se afasta significativamente da linha azul (maus), especialmente entre as probabilidades de 0,1 e 0,3, indicando que o modelo de Regressão Logística é eficaz em identificar bons pagadores. A distância máxima entre as duas curvas, conhecida como estatística KS, é um indicativo da eficácia do modelo na discriminação entre classes positivas e negativas.

Gradient Boosting

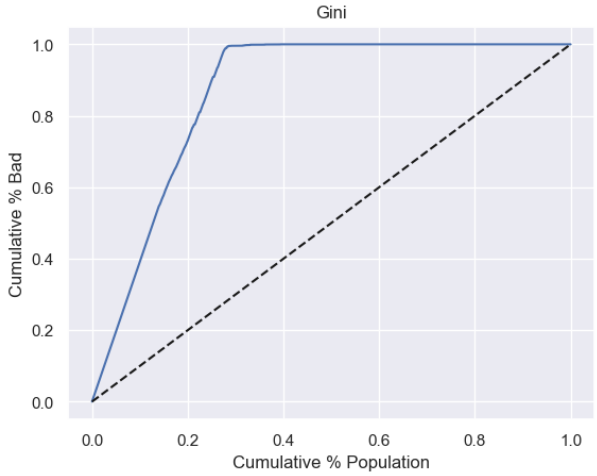

Figura 4.12: Gini Gradient Boost

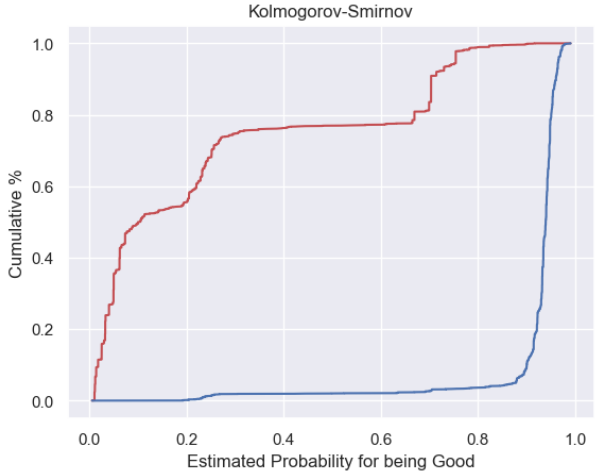

Figura 4.13: Kolmogorov Gradient Boost

Os gráficos de Gini e Kolmogorov-Smirnov para o modelo Gradient Boosting, nas Figuras 4.12 e 4.13, indicam o desempenho do modelo na separação das classes. A curva Gini do modelo de Gradient Boosting também mostra uma excelente capacidade discriminatória. Similar à Regressão Logística, a curva se afasta da linha de igualdade de maneira significativa nos primeiros 20% da população. Este modelo é igualmente eficiente em concentrar a maioria dos inadimplentes em uma pequena parte da população, o que é crucial para a análise de crédito. A área grande entre a curva e a linha diagonal indica que o Gradient Boosting é altamente eficaz em separar as classes.

A curva Kolmogorov-Smirnov do modelo de Gradient Boosting também demonstra uma excelente capacidade discriminatória. A linha vermelha se afasta ainda mais da linha

azul em comparação com a Regressão Logística, especialmente entre as probabilidades de 0,1 e 0,4. Isso sugere que o modelo de Gradient Boosting tem uma capacidade ainda maior de discriminar entre bons e maus pagadores, tornando-o altamente eficaz para a análise de crédito.

Random Forest

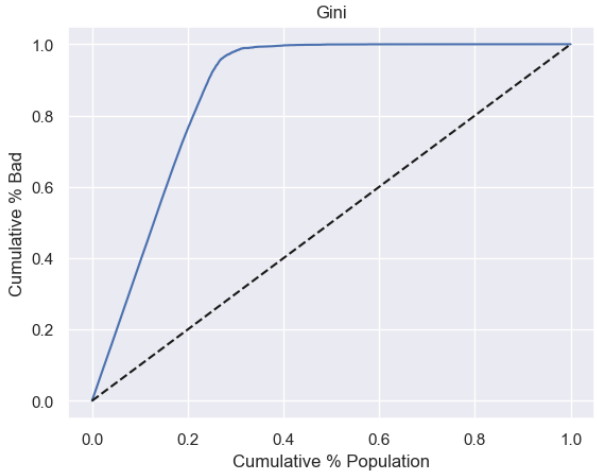

Figura 4.14: Gini Random Forest

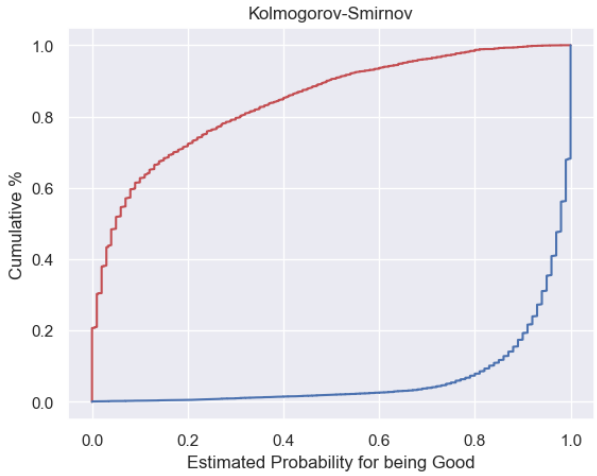

Figura 4.15: Kolmogorov Random Forest

As Figuras 4.14 e 4.15 exibem os gráficos de Gini e Kolmogorov-Smirnov para o modelo Random Forest. A curva Gini do modelo de Random Forest revela uma excelente capacidade discriminatória, semelhante aos modelos de Regressão Logística e Gradient Boosting. A curva se afasta da linha de igualdade rapidamente nos primeiros 20% da população, mostrando que o modelo é muito eficiente em identificar os "maus"na população

analisada. A área substancial entre a curva e a linha diagonal sugere que o coeficiente de Gini é alto, indicando a eficácia do Random Forest na discriminação de classes. A curva KS do modelo de Random Forest revela uma excelente capacidade discriminatória, similar à do Gradient Boosting. A linha vermelha se afasta significativamente da linha azul, especialmente entre as probabilidades de 0,1 e 0,3, indicando que o modelo de Random Forest é muito eficiente em identificar bons pagadores. A distância significativa entre as duas curvas sugere um alto valor de estatística KS, confirmando a eficácia do modelo na discriminação de classes.

Árvore de Decisão

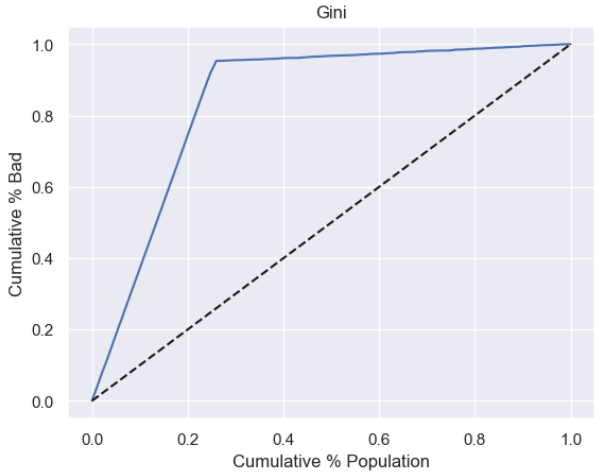

Figura 4.16: Gini Decision Tree

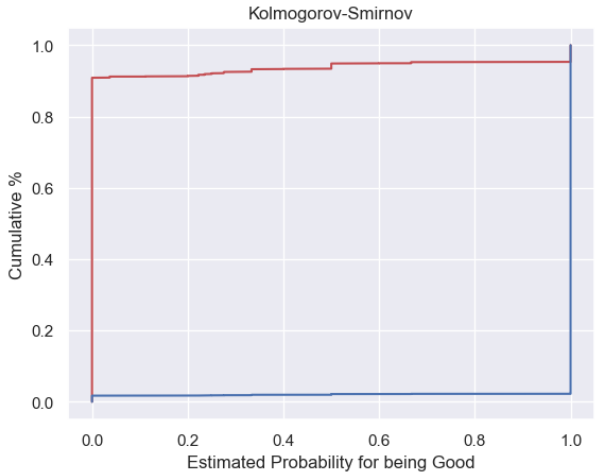

Figura 4.17: Kolmogorov Decision Tree

Os gráficos de Gini e Kolmogorov-Smirnov para o modelo Árvore de Decisão, apre-

sentados nas Figuras 4.16 e 4.17, fornecem uma visão detalhada sobre o desempenho do modelo. A curva Gini do modelo de Decision Tree apresenta uma boa capacidade discriminatória, embora seja ligeiramente inferior em comparação com os outros três modelos. A curva se afasta da linha de igualdade, mas com um desempenho um pouco menor nos primeiros 20% da população. Isso indica que, embora o modelo de Decision Tree seja eficiente em identificar inadimplentes, ele não é tão eficaz quanto os modelos de Regressão Logística, Gradient Boosting e Random Forest. A área entre a curva e a linha diagonal é menor, sugerindo um coeficiente de Gini ligeiramente inferior.

A curva Kolmogorov-Smirnov do modelo de Decision Tree apresenta uma boa capacidade discriminatória. A linha vermelha se afasta significativamente da linha azul ao longo de toda a faixa de probabilidades, indicando que o modelo é eficaz em separar os bons dos maus pagadores. No entanto, a separação não é tão acentuada quanto nos modelos de Gradient Boosting e Random Forest, sugerindo uma discriminação ligeiramente inferior.

4.3.3 Discussões das métricas de desempenho.

Os resultados indicam que a Regressão Logística, apesar de sua simplicidade, oferece um desempenho robusto em termos de AUC ROC, F1-Score e Acurácia. O Random Forest e o Gradient Boosting também demonstram desempenhos excepcionais, especialmente no recall, o que pode ser vantajoso em aplicações onde a detecção de falsos negativos é crítica. A Árvore de Decisão, embora apresente um desempenho competitivo em termos de acurácia e F1-Score, é superada pelos outros modelos em termos de AUC ROC.

Ao comparar as curvas Gini dos quatro modelos, podemos observar que os modelos de Regressão Logística, Gradient Boosting e Random Forest apresentam desempenhos muito semelhantes e superiores ao modelo de Decision Tree. Esses três modelos possuem curvas Gini que se afastam significativamente da linha de igualdade nos primeiros 20% da população, indicando uma alta capacidade de discriminação e eficácia na identificação de inadimplentes.

O modelo de Decision Tree, embora eficaz, apresenta um desempenho ligeiramente inferior, com uma área menor entre a curva e a linha diagonal. Isso sugere que, enquanto a Decision Tree é uma ferramenta útil, ela pode não ser tão robusta quanto os outros

modelos na discriminação de classes em um contexto de análise de crédito.

A análise das curvas Gini confirma que os modelos de Regressão Logística, Gradient Boosting e Random Forest são altamente eficazes e confiáveis para a análise de crédito, proporcionando uma ferramenta robusta para a identificação de inadimplentes. O modelo de Decision Tree, embora eficiente, é ligeiramente menos eficaz comparado aos outros modelos. A escolha do modelo ideal deve considerar o contexto específico da aplicação e a importância relativa de cada métrica de desempenho. Em aplicações onde a identificação precisa de clientes de alto risco é crucial, os modelos de Regressão Logística, Gradient Boosting e Random Forest são recomendados.

Ao comparar as curvas KS dos quatro modelos, observamos que todos apresentam uma boa capacidade discriminatória. No entanto, os modelos de Gradient Boosting e Random Forest mostram uma separação mais clara entre as distribuições cumulativas dos bons e maus em comparação com a Regressão Logística e a Decision Tree, indicando um desempenho ligeiramente superior.

A análise das curvas Kolmogorov-Smirnov confirma que os modelos de Gradient Boosting e Random Forest possuem uma excelente capacidade de discriminação entre bons e maus pagadores, com uma ligeira vantagem sobre a Regressão Logística e a Decision Tree. A escolha do modelo ideal deve considerar o contexto específico da aplicação e a importância relativa de cada métrica de desempenho. Em aplicações onde a identificação precisa de clientes de alto risco é crucial, os modelos de Gradient Boosting e Random Forest são altamente recomendados.

Em suma, a análise comparativa das diferentes métricas de desempenho (AUC ROC, F1-Score, Acurácia, Curvas Gini e Curvas Kolmogorov-Smirnov) demonstra que os modelos de Gradient Boosting e Random Forest possuem a melhor capacidade discriminatória, seguidos de perto pela Regressão Logística. O modelo de Decision Tree, embora útil, apresenta desempenho inferior comparado aos outros modelos.

Com base nesses resultados, a utilização dos modelos de Gradient Boosting e Random Forest para aplicações de análise de crédito, especialmente em cenários onde a identificação precisa de clientes de alto risco é essencial. A Regressão Logística também é uma opção muito viável, particularmente devido à sua simplicidade e robustez.

4.4 Análise da Importância das Features para os modelos de predição

A análise de importâncias das variáveis permite identificar quais variáveis têm maior influência na predição da variável alvo. Foi realizada a análise comparativa das importâncias das variáveis dos três modelos: Gradient Boosting, Random Forest e Decision Tree e a análise dos coeficientes da regressão logística. A análise identifica as características mais relevantes para cada modelo e aquelas que não apresentam relevância significativa.

4.4.1 Importâncias das Features por Modelo

Feature Name	**Importância**
region:northeast	0,34
region:southeast	0,12
age_group:48,00 - 56,00	0,12
job_count:0	0,13
region:south	0,07
marital_status:married	0,03
region:midwest	0,04
betweenness_centrality:(-0,232, 33,149]	0,001
closeness_centrality:-0,00 - 0,11	0,001
community_louvain:64,00 - 71,00	0,0006

Tabela 4.3: Top 10 Características por Importância (Gradient Boost)

A tabela 4.3, apresenta as variáveis mais importantes para o Gradient Boosting.

Feature Name	**Importância**
region:northeast	0,31
job_count:0	0,09
age_group:48,00 - 56,00	0,08
region:southeast	0,003
region:south	0,01
marital_status:married	0,03
marital_status:single	0,002
betweenness_centrality:99,447 - 132,597	0,001
betweenness_centrality:132,597 - 165,746	0,001
closeness_centrality:-0,00 - 0,11	0,0008

Tabela 4.4: Top 10 Características por Importância (Random Forest)

A tabela 4.4, apresenta as variáveis mais importantes para o Random Forest

Feature Name	**Importância**
region:northeast	0,12
region:southeast	0,08
job_count:0	0,06
age_group:48,00 - 56,00	0,07
region:south	0,03
marital_status:married	0,02
marital_status:single	0,01
betweenness_centrality:99,447 - 132,597	0,002
betweenness_centrality:132,597 - 165,746	0,002
community_louvain:64,00 - 71,00	0,0006

Tabela 4.5: Top 10 Características por Importância (Decision Tree)

A tabela 4.5, apresenta as variáveis mais importantes para Decision Tree.

4.4.2 Análise dos Coeficientes da Regressão Logística

Os coeficientes da regressão logística indicam a força e a direção da relação de cada característica com a variável dependente. Coeficientes positivos sugerem uma associação positiva, onde o aumento na característica está relacionado a um aumento na probabilidade do resultado alvo. Coeficientes negativos indicam uma associação negativa, onde o aumento na característica está relacionado a uma diminuição na probabilidade do resultado alvo. As Tabelas 4.6 e 4.7 mostram as características com os coeficientes mais altos e mais baixos, respectivamente.

Feature Name	**Coeficiente**
marital_status:married	9,64
marital_status:single	8,43
age_group:17,00 - 23,00	8,04
age_group:23,00 - 25,00	8,13
age_group:25,00 - 26,00	8,01
age_group:26,00 - 28,00	8,24
age_group:28,00 - 29,00	8,10
age_group:29,00 - 31,00	9,01
age_group:31,00 - 32,00	9,46
age_group:32,00 - 34,00	9,63

Tabela 4.6: Top 10 Características por Coeficiente (Mais Altos)

Feature Name	Coeficiente
job_count:0	-4,67
betweenness_centrality:(-0,232, 33,149]	-2,02
betweenness_centrality:(33,149, 66,298]	-0,87
closeness_centrality:-0,00 - 0,11	-1,05
closeness_centrality:0,11 - 1,13	-0,67
closeness_centrality:1,13 - 4,01	-0,52
degree_centrality:0,06 - 0,12	-1,59
degree_centrality:0,12 - 226,59	-1,73
eigenvector_centrality:-0,00 - 0,00	-0,73
eigenvector_centrality:0,00 - 0,00	-1,14

Tabela 4.7: Top 10 Características por Coeficiente (Mais Baixos)

4.4.3 Importâncias das Características nos Três Modelos

A Tabela 4.8 mostra as características consideradas importantes em mais de um modelo.

Feature Name	Modelos	Importância
region:northeast	Gradient Boost, Random Forest, Decision Tree	Alta
job_count:0	Gradient Boost, Random Forest, Decision Tree	Alta
region:southeast	Gradient Boost, Random Forest, Decision Tree	Alta
age_group:48,00 - 56,00	Gradient Boost, Random Forest, Decision Tree	Alta
region:south	Gradient Boost, Random Forest, Decision Tree	Alta
marital_status:married	Gradient Boost, Decision Tree	Moderada
marital_status:single	Random Forest, Decision Tree	Moderada

Tabela 4.8: Características Importantes em Mais de um Modelo

4.4.4 Características sem Relevância preditiva

As características que não apresentaram relevância em nenhum dos modelos analisados: age_group:56,00, marital_status:divorce, gender:male, iscapitalcity:False, current_state:amazonas hometown_state_region:norte, region:north, language_count:poliglota, contact_info:Fals, nickname:True, foreigner:True, education:fundamental, following_count:False, friends_link:0, friend_count:2 ,job_count:1

4.4.5 Discussão da importância das variaveis.

A análise das importâncias das variáveis nos quatro modelos de aprendizado de máquina (Gradient Boosting, Random Forest, Decision Tree e Regressão Logística) revelou algumas características comuns que são altamente relevantes em todos os modelos. Características como *region:northeast*, *job_count:0*, *region:southeast*, *age_group:48,00 - 56,00* e *region:south* se destacaram consistentemente entre as mais importantes. Isso sugere que essas características têm uma forte influência na predição da variável alvo e devem ser consideradas críticas no desenvolvimento de modelos preditivos.

Além disso, a análise dos coeficientes da regressão logística indicou que características como *marital_status:married* e várias faixas etárias (*age_group:17,00 - 34,00*) têm coeficientes positivos altos, indicando uma associação positiva significativa com a variável dependente. Por outro lado, características como *job_count:0* e diferentes intervalos de *betweenness_centrality* apresentaram coeficientes negativos, sugerindo uma associação negativa com a variável dependente.

Outro ponto importante é a identificação de características que não apresentaram relevância em nenhum dos quatro modelos analisados. Características como *age_group:56,00 - 117,00*, *marital_status:divorced*, *gender:male*, *iscapitalcity:False* e *current_state:amazonas* tiveram importâncias nulas em todos os modelos. Isso indica que essas características podem ser irrelevantes para a predição e podem ser removidas do modelo para simplificação e melhor desempenho.

Para essas características irrelevantes, há uma necessidade clara de realizar engenharia de variáveis adicional para torná-las mais úteis para a modelagem. Técnicas como criação de novas interações entre variáveis, transformação de dados, e uso de métodos avançados

de seleção de variáveis podem ser aplicadas para tentar extrair valor dessas características inicialmente irrelevantes. Além disso, considerar a coleta de dados adicionais ou fontes de dados alternativas pode ajudar a aumentar a relevância dessas características.

A consistência das importâncias das variáveis nos quatro modelos reforça a robustez dessas características na predição. No entanto, a análise também destaca a necessidade de considerar a especificidade de cada modelo, já que algumas características apresentam importâncias variadas entre os modelos. Esta percepção pode ser útil para ajustar e otimizar os modelos de aprendizado de máquina, considerando as características mais e menos relevantes identificadas nesta análise.

Em conclusão, a análise das importâncias das variáveis e dos coeficientes dos quatro modelos fornece uma compreensão detalhada das variáveis mais influentes na predição da variável alvo. Isso pode guiar futuras otimizações de modelos, bem como informar decisões sobre a inclusão ou exclusão de características no processo de modelagem.

4.5 Resultados das Regras de Corte

Para determinar o threshold ideal, utilizamos duas abordagens: a Curva ROC e a Precision-Recall.

4.5.1 Curva ROC e AUC

A Curva ROC (Receiver Operating Characteristic) plota a taxa de verdadeiros bons (TPR) contra a taxa de falsos bons (FPR) para diferentes thresholds. A Área Sob a Curva (AUC) é uma medida de quão bem o modelo separa as classes. Um valor de AUC mais próximo de 1 indica um modelo melhor. O ponto ideal no gráfico ROC é aquele mais próximo do canto superior esquerdo (onde TPR é alto e FPR é baixo).

Figura 4.18: ROC régua de corte regressão logística

ROC Curve: A curva ROC 4.18 mostra que o threshold de 0,51 resulta em uma linha diagonal, indicando que o modelo tem um desempenho próximo ao de uma classificação aleatória (AUC = 0,52). Isso sugere que este threshold pode não ser o mais eficiente para discriminar entre bons e maus pagadores.

4.5.2 Precision-Recall e F1-Score

A curva Precision-Recall mostra as relações entre precisão, recall e F1-score para diferentes thresholds. O threshold ideal para maximizar o F1-score, sendo a média harmônica da precisão e do recall, é aproximadamente 0,998.

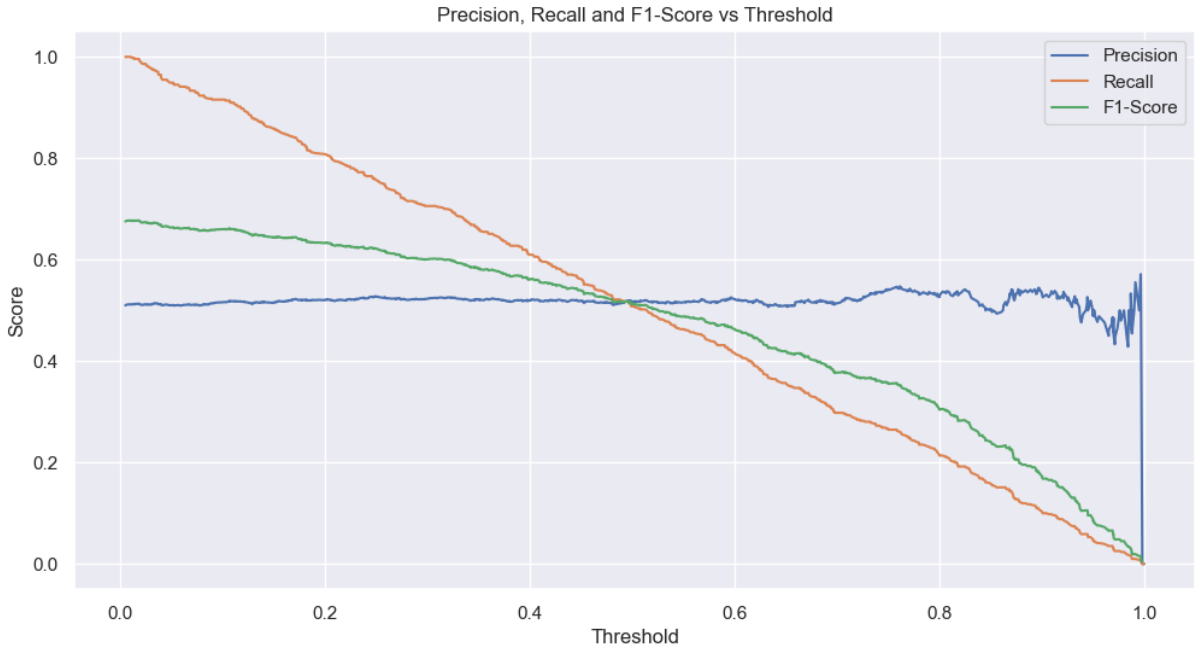

Figura 4.19: Recall e F1-score vs Threshold

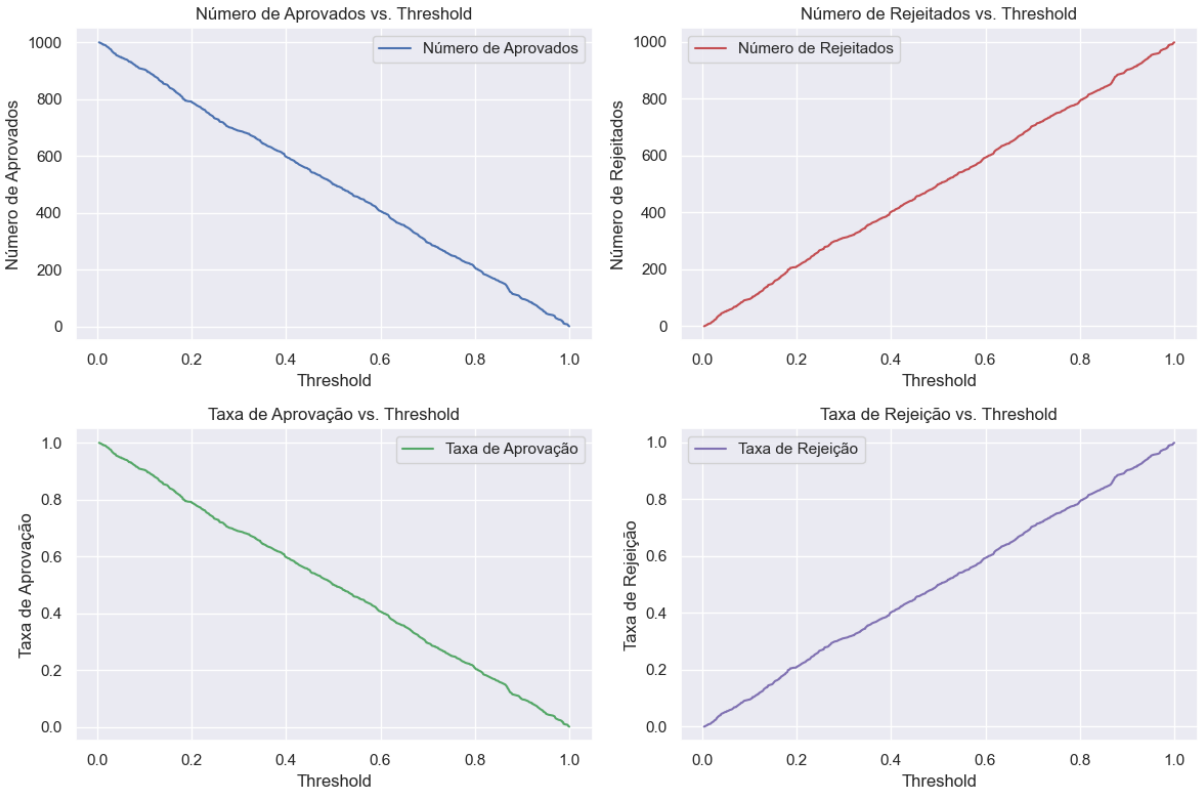

Figura 4.20: Taxas de aprovação e rejeição vs Threshold

Conforme 4.20:

- **Número de Aprovados vs. Threshold**: À medida que o threshold aumenta, o número de aprovados diminui linearmente. No threshold ideal (0,51), o número de aprovados é moderado, indicando um equilíbrio entre aprovação e rejeição.

- **Número de Rejeitados vs. Threshold**: Similarmente, o número de rejeitados aumenta linearmente com o aumento do threshold. No threshold ideal, o número de rejeitados é moderado.

- **Taxa de Aprovação vs. Threshold**: A taxa de aprovação diminui conforme o threshold aumenta. No threshold de 0,51, a taxa de aprovação está aproximadamente na metade, refletindo um equilíbrio entre aceitação e rejeição.

- **Taxa de Rejeição vs. Threshold**: A taxa de rejeição aumenta com o aumento do threshold. No threshold ideal, a taxa de rejeição está próxima de 0,5, indicando um equilíbrio entre rejeições e aprovações.

4.5.3 Discussão dos Thresholds Ideais

- **Threshold Ideal (ROC)**: 0,51, - Esse threshold é escolhido para equilibrar a taxa de verdadeiros bons (TPR) e a taxa de falsos bons (FPR). - É útil ao se querer um bom trade-off entre detectar o maior número possível de bons verdadeiros e minimizar os falsos bons.

- **Threshold Ideal (Precision-Recall)**: 0,997 - Esse threshold maximiza o F1-score, sendo a média harmônica da precisão e do recall. - É útil ao se desejar um equilíbrio entre a precisão (a proporção de previsões positivas que são realmente positivas) e o recall (a proporção de reais bons, corretamente identificados).

O gráfico "Precision, Recall e F1-Score vs. Threshold"4.19 sugere que um threshold em torno de 0,4 a 0,5 pode oferecer um bom equilíbrio entre precisão e recall, resultando em um F1-score robusto.

Capítulo 5

CONCLUSÃO E TRABALHOS FUTUROS

A presente pesquisa investigou o desenvolvimento de um modelo de análise de crédito baseado em dados de redes sociais, visando explorar como essas informações podem ser integradas ou substituir as análises de crédito tradicionais através do comprimento dos 5 C's do crédito: Caráter, Capacidade, Capital, Colateral e Condições. Através da coleta e análise de dados de redes sociais, identificamos diversas variáveis que contribuem para a avaliação de cada pilar, ao mesmo tempo, em que reconhecemos as limitações e propondo soluções para mitigá-las.

O pilar do Caráter foi significativamente enriquecido pelo uso de dados de redes sociais. Campos como *"friend_count"*, *"follower_count"*, *"following_count"*, *"posts"*, *"education"*, *"employed"*, *"marital_status"* e *"language_count"* foram utilizados para construir uma visão detalhada do comportamento e da confiabilidade do indivíduo. A análise das interações sociais, frequência de postagens e conteúdo compartilhado forneceu percepções valiosas sobre a responsabilidade e a estabilidade emocional do solicitante, permitindo uma avaliação mais precisa do seu caráter.

Para avaliar a Capacidade de pagamento, utilizamos dados como *"employed"*, *"job_count"*, *"jobs"*, *"education"*, *"friend_count"* e *"friends_link"*. Esses campos forneceram informações sobre a estabilidade profissional e a rede de suporte social do indivíduo. A combinação desses dados permite inferir a capacidade de geração de renda contínua e a probabilidade

de apoio financeiro em momentos de necessidade. No entanto, reconhecemos a limitação de não possuir dados financeiros concretos, como renda e despesas mensais.

O pilar do Capital não pôde ser completamente atendido apenas com dados de redes sociais. Embora tenhamos utilizado campos como *"posts"* e *"photos"* para identificar sinais de posses significativas e estilo de vida, esses dados não substituem informações financeiras precisas, como saldos bancários e investimentos. Para mitigar essa deficiência, sugerimos integrar dados financeiros tradicionais, como extratos bancários e declarações de patrimônio, para complementar as inferências feitas a partir das redes sociais.

A identificação de Colateral foi outro desafio, pois dados como *"posts"*, *"photos"*, *"current_city"*, *"current_state"*, *"hometown"* e *"hometown_state"* fornecem apenas indícios indiretos sobre ativos tangíveis. A ausência de detalhes precisos sobre imóveis ou veículos impede uma avaliação completa deste pilar. Para mitigar essa limitação, recomendamos solicitar documentos comprobatórios de propriedade, como escrituras de imóveis e registros de veículos, para garantir uma avaliação robusta de colaterais.

As Condições econômicas gerais e específicas do setor foram parcialmente abordadas mediante campos como *"region"*, *"current_state_region"*, *"political_statement"*, *"nationality"*, *"interests"* e *"iscapitalcity"*. Esses dados fornecem contexto sobre a estabilidade econômica e política do ambiente onde o indivíduo vive. No entanto, para uma análise mais precisa, é necessário integrar dados macroeconômicos e setoriais obtidos de fontes externas, como relatórios econômicos e indicadores de mercado.

A análise pode ser aprimorada com a inclusão de dados financeiros diretos, como histórico de crédito e ativos tangíveis, para complementar as proxies utilizadas. Dados Econômicos Regionais Detalhados: Dados econômicos regionais detalhados podem fornecer uma visão mais precisa das condições econômicas enfrentadas pelos indivíduos.

Informações adicionais sobre o comportamento financeiro, como padrões de gasto e poupança, poderiam melhorar a avaliação do caráter e da capacidade dos tomadores de crédito. Desempenho Preditivo dos Modelos Os modelos de aprendizado de máquina testados (Regressão Logística, Árvore de Decisão, Floresta Randômica e Gradient Boosting) mostraram desempenhos robustos, com a Regressão Logística, Gradient Boosting e Random Forest se destacando. A análise de importância das variáveis indicou que variáveis geográficas, estado civil, idade e centralidade em redes sociais são fatores significativos

para a predição de inadimplência.

Resposta às Perguntas de Pesquisa 1. Quais padrões e comportamentos financeiros podem ser identificados em dados de redes sociais e como eles podem ser integrados aos modelos tradicionais de análise de crédito? Os dados de redes sociais permitiram identificar vários padrões e comportamentos financeiros relevantes:

Educação e Emprego: Informações sobre o nível de educação e histórico de emprego (número de empregos e posições ocupadas) revelaram-se indicadores importantes da capacidade financeira dos indivíduos. Pessoas com maior nível educacional e histórico profissional estável mostraram maior capacidade de geração de renda.

Interações Sociais: A quantidade de amigos, seguidores e o nível de engajamento nas redes sociais indicaram o suporte social e o capital social, relevantes para avaliar o capital do indivíduo.

Comportamento e Estilo de Vida: Análises de postagens, fotos e atividades nas redes sociais forneceram percepções sobre o estilo de vida dos indivíduos, suas responsabilidades e estabilidade emocional, influenciando a avaliação de caráter.

Esses dados podem ser integrados aos modelos tradicionais de análise de crédito para complementar as informações financeiras convencionais, proporcionando uma visão mais holística do perfil de crédito do indivíduo.

2. Quais são as vantagens que a análise de dados de redes sociais traz para a análise de crédito? A análise de dados de redes sociais oferece várias vantagens:

As redes sociais fornecem uma vasta quantidade de dados em tempo real que podem ser usados para complementar os dados financeiros tradicionais.

A análise de interações e comportamentos nas redes sociais oferece percepções valiosos sobre a personalidade e estabilidade emocional dos indivíduos, aspectos que são difíceis de capturar apenas com dados financeiros.

Suporte Social e Capital Social: As redes sociais revelam a rede de suporte e o capital social dos indivíduos, que são indicadores importantes de sua capacidade de enfrentar dificuldades financeiras.

Atualização Constante: Diferente dos dados financeiros tradicionais, que podem ser

estáticos e desatualizados, os dados de redes sociais são dinâmicos e refletem mudanças na vida dos indivíduos em tempo real.

3. É viável substituir a análise de crédito baseada em dados financeiros tradicionais por uma baseada em dados de redes sociais? Embora os dados de redes sociais ofereçam vantagens significativas, a substituição completa da análise tradicional não é viável por várias razões:

Dados Financeiros Concretos: Informações financeiras concretas, como histórico de crédito, renda e ativos, são fundamentais para uma avaliação precisa e completa do risco de crédito.

Limitações dos Dados Sociais: Dados de redes sociais podem ser incompletos ou imprecisos, e não cobrem todos os aspectos necessários para uma avaliação de crédito abrangente.

Complementaridade: A melhor abordagem é a integração de dados de redes sociais com dados financeiros tradicionais. Esta combinação oferece uma visão mais rica e detalhada do perfil de crédito dos indivíduos, melhorando a precisão e eficácia da análise de crédito.

Como trabalhos futuros recomenda-se a integração contínua de dados financeiros tradicionais com dados de redes sociais para uma análise de crédito mais completa e precisa.

Aprimoramento de Variáveis: Investir na coleta de dados econômicos regionais mais detalhados e histórico de crédito para complementar as variáveis de redes sociais.

Exploração de Novas Fontes de Dados: Ampliar a pesquisa para incluir outras plataformas de redes sociais, como LinkedIn e Instagram, e novas métricas de interação social.

Análise Longitudinal: Realizar análises longitudinais para entender como as variáveis de redes sociais evoluem ao longo do tempo e seu impacto na análise de crédito.

Integração com Dados Auxiliares: Incorporar dados auxiliares de fontes como o IBGE e pesquisas de salários por profissão (por exemplo, Love Mondays) para enriquecer a análise das condições econômicas e capacidades financeiras dos indivíduos. Em conclusão, as variáveis analisadas mostraram-se na maioria alinhadas com os cinco pilares tradicionais da avaliação de crédito. No entanto, há espaço para a inclusão de dados adicionais que

poderiam aprimorar ainda mais a precisão e a robustez da análise. A utilização de dados de redes sociais na análise de crédito se mostrou promissora, proporcionando percepções valiosas e complementando as abordagens tradicionais. No entanto, a combinação de ambos os tipos de dados oferece a melhor estratégia para uma análise de crédito mais precisa e eficaz, contribuindo para a melhoria das práticas de concessão de crédito e gestão de risco financeiro.

Referências Bibliográficas

ADAMS, R. *Economic Conditions and Credit Risk*. Washington: Brookings Institution Press, 2017.

ALENCAR, N. d. A. *Título do Livro*. Cidade: Editora, 2000. 61-63 p.

BAIDEN, J. E. The 5 c's of credit in the lending industry. *SSRN*, 2011. Disponível em: <https://ssrn.com/abstract=1872804>.

BAIDEN, J. E. The 5 c's of credit in the lending industry. *Journal of Banking & Finance*, v. 35, n. 3, p. 145–156, 2011.

Banco Central do Brasil. *Relatório de Estabilidade Financeira*. 2023. Acesso em: 2024-06-06. Disponível em: <https://www.bcb.gov.br/relatorio/estabilidade-financeira>.

BAZARBASH, M. Fintech in financial inclusion: Machine learning applications in assessing credit risk. *FinPlanRN: Other Finance Planning Fundamentals (Topic)*, 2019.

BLONDEL, V. D. et al. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, v. 2008, p. P10008, 2008.

Boa Vista. *Pesquisa de inadimplência e endividamento*. 2022. Disponível em: https://www.boavistaservicos.com.br. Acesso em: 10 jun. 2024.

BOYD, D. M.; ELLISON, N. B. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, v. 13, n. 1, p. 210–230, 2007.

BREIMAN, L. Random forests. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001.

BROWN, M. *Assessing Capital in Credit Analysis*. Boston: Harvard Business Review Press, 2018.

CAIANI, M.; PARENTI, R. Social network analysis in the study of terrorism and radicalism: from theory to practice. *International Journal of Social Research Methodology*, v. 14, n. 2, p. 113–131, 2011. Disponível em: <https://www.redalyc.org-/articulo.oa?id=>.

CAOUETTE, J. B.; ALTMAN, E. I.; NARAYANAN, P. *Managing Credit Risk: The Next Great Financial Challenge*. New York: John Wiley & Sons, 1998.

CFA Institute. *Fundamentals of Credit Analysis*. 2024. Acesso em: 2024-06-06. Disponível em: <https://www.cfainstitute.org/en/membership/professional-development/refresher-readings/fundamentals-credit-analysis>.

Chartered Financial Analyst Institute. *Fundamentals of Credit Analysis*. 2023. Acesso em: 2023-12-06. Disponível em: <https://www.cfainstitute.org/en/membership-/professional-development/refresher-readings/fundamentals-credit-analysis>.

CLARK, K. *Character Assessment through Social Media*. Berlin: Springer, 2018.

CNUDDE, S. D. et al. What does your facebook profile reveal about your creditworthiness? using alternative data for microfinance. *Journal of the Operational Research Society*, Springer, v. 70, n. 3, p. 353–363, 2019.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine Learning*, v. 20, n. 3, p. 273–297, 1995.

COX, D. R. *The Analysis of Binary Data*. [S.l.]: CRC Press, 1989.

CROUHY, M.; GALAI, D.; MARK, R. *Risk Management*. [S.l.]: McGraw-Hill, 2001.

CROUHY, M.; GALAI, D.; MARK, R. *Risk Management*. New York: McGraw-Hill, 2001.

Código de Defesa do Consumidor. *Lei nº 8.078/90 - Lei de Proteção ao Consumidor*. 1990. Acesso em: 2024-06-06. Disponível em: <https://www.planalto.gov.br/ccivil_03-/leis/l8078compilado.htm>.

DAVIS, J. *Collateral Valuation and Social Media Insights*. Los Angeles: Sage Publications, 2020.

EMIRBAYER, M. Network analysis: An analytical strategy for investigating social structures. *Sociological Theory*, v. 12, n. 1, p. 141–152, 1994. Disponível em: <https://www.scielo.br/j/cies/article/view>.

EQUIFAX. *Equifax Credit Score: Understanding Your Credit Score*. 2021. Disponível em: https://www.equifax.com. Acesso em: 10 jun. 2024.

EXPERIAN, S. *Estratégias de Recuperação de Crédito*. São Paulo: Serasa Experian, 2019.

EXPERIAN, S. *Serasa Experian Credit Score: Understanding Your Credit Score*. 2022. Disponível em: https://www.serasaexperian.com.br. Acesso em: 10 jun. 2024.

EXPERIAN, S. *Quais informações compõem o Serasa Score?* 2024. Acesso em: 10 jun. 2024. Disponível em: <https://ajuda.pmedigital.serasaexperian.com.br/hc/pt-br-/articles/4485105496979-Quais-informa%C3%A7%C3%B5es-comp%C3%B5em-o-Serasa-Score>.

FINANçAS, R. de. Eficácia da cobrança amigável na recuperação de dívidas. *Revista de Finanças*, v. 45, p. 123–145, 2020.

FISHER, R. A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, v. 7, n. 2, p. 179–188, 1936.

FORTUNATO, S. Community detection in graphs. *Physics Reports*, v. 486, p. 75–174, 2010.

FRANKLIN, J. Ethical considerations of web scraping. *Journal of Information Ethics*, v. 29, n. 2, p. 123–135, 2020.

FRANKLIN, P. *Web Scraping com R*. 2020. Disponível em: `https://rpubs.com/pedrofranklin/web-scraping`. Acesso em: 2024-06-07.

GIL, A. C. *Métodos e técnicas de pesquisa social*. 6. ed. São Paulo: Atlas, 2006.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016.

GROSSO, M. Eficácia da negociação amigável na cobrança de dívidas. *Revista de Administração de Empresas*, v. 61, n. 4, p. 321–333, 2021.

GROVER, A.; LESKOVEC, J. node2vec: Scalable feature learning for networks. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, ACM, San Francisco, p. 855–864, 2016.

HAND, D. J.; HENLEY, W. E. Modelling credit scoring. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Wiley Online Library, v. 48, n. 4, p. 525–541, 2001.

HAND, D. J.; HENLEY, W. E. Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, v. 160, n. 3, p. 523–541, 2001.

HARVARD. The psychology of debt collection: Using empathy and personalization. *Harvard Business Review*, 2018.

HAWE, P.; WEBSTER, C.; SHIELL, A. A glossary of terms for navigating the field of social network analysis. *Journal of Epidemiology and Community Health*, v. 58, n. 12, p. 971–975, 2004. Disponível em: <https://www.redalyc.org/articulo.oa?id=>.

HOSMER, D. W.; LEMESHOW, S.; STURDIVANT, R. X. *Applied Logistic Regression*. [S.l.]: John Wiley & Sons, 2013.

IBGE. *Desemprego no Brasil*. 2020. Acesso em: 2024-06-06. Disponível em: <https://www.ibge.gov.br/indicadores/desemprego>.

JOHNSON, E. *Social Media and Financial Decision-Making*. London: Routledge, 2019.

KEPIOS, I. *Facebook Users, Stats, Data and Trends*. 2023. Disponível em: `https://datareportal.com/essential-facebook-stats`. Acesso em: 12/12/2023.

KIM, M.; SOHN, S. Y. Real-time credit monitoring using financial transaction data. *Journal of Financial Services Research*, Springer, v. 57, n. 2, p. 169–190, 2020.

KOCH, T. W.; MACDONALD, S. S. *Bank Management*. Fort Worth: Dryden Press, 2000.

KUMAR, A.; GUPTA, A. Predictive models for credit default: Logistic regression vs random forest. *Journal of Financial Services Marketing*, v. 25, n. 3, p. 45–55, 2020.

KUMAR, S.; SINHA, R. Social network-based models for credit risk assessment. *Journal of Financial Services Marketing*, v. 25, n. 2, p. 89–101, 2020.

LAZEGA, E.; HIGGINS, J. P. Networking in context: Explaining organizational behavior through social network analysis. *Sociological Review*, v. 62, n. 1, p. 34–48, 2014. Disponível em: <https://www.redalyc.org/articulo.oa?id=>.

LEE, S.; CHOI, J. Predictive modeling using social network data for credit risk assessment. *Journal of Credit Risk Management*, v. 17, n. 1, p. 78–94, 2020.

LEE, S.; KIM, H. Comparative study on credit scoring models. *Journal of Credit Risk Management*, v. 18, n. 2, p. 101–115, 2020.

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.

MDPI. Framework for social media analysis based on hashtag research. *Applied Sciences*, 2024. Disponível em: <https://www.mdpi.com>.

META. *Política de Dados*. 2023. Disponível em: `https://www.facebook.com/policy.php`. Acesso em: 2024-06-07.

MICHEL, J. *Pesquisa de mercado: metodologia e prática*. 2. ed. [S.l.]: Atlas, 2005.

MICHEL, M. H. Metodologia e pesquisa científica em ciências sociais. *São Paulo: Atlas*, 2005.

MILLER, T. *Collateral in Credit Markets*. Cambridge: Cambridge University Press, 2016.

NEON. *Birôs de crédito: conheça os principais e o que fazem*. 2024. Acesso em: 12 jan. 2024. Disponível em: <https://neon.com.br/aprenda/economizar-dinheiro/biros-de-credito/>.

NEWMAN, M. *Networks: An Introduction*. Oxford: Oxford University Press, 2010. ISBN 9780199206650.

PEROZZI, B.; AL-RFOU, R.; SKIENA, S. Deepwalk: Online learning of social representations. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*, ACM, New York, p. 701–710, 2014.

QUINLAN, J. R. Induction of decision trees. *Machine Learning*, v. 1, n. 1, p. 81–106, 1986.

QUOD. *Consulta gratuita de débitos e informações cadastrais*. 2023. Acesso em: 2024-06-07. Disponível em: <https://www.quod.com.br/>.

SAUNDERS, A.; CORNETT, M. M. *Financial Institutions Management: A Risk Management Approach*. New York: McGraw-Hill/Irwin, 2007.

SCIENCEDIRECT. Analytics of social media data – state of characteristics and future directions. *ScienceDirect*, 2024. Disponível em: <https://www.sciencedirect.com>.

SCPC, B. V. *Relatório Anual de Inadimplência*. São Paulo: Boa Vista Serviços, 2020.

SCPC, B. V. *Pesquisa de inadimplência e endividamento*. 2022. Disponível em: https://www.boavistaservicos.com.br. Acesso em: 10 jun. 2024.

SCPC, S. d. P. a. C. Estratégias de cobrança amigável e seu impacto na inadimplência. *Revista de Crédito e Cobrança*, v. 32, p. 67–89, 2020.

Serasa. *Análise de crédito: o que é, para que serve e como é feita*. 2023. Acesso em: 2023-07-06. Disponível em: <https://www.serasa.com.br/credito/blog/analise-de-credito-o-que-e-para-que-serve-e-como-e-feita/>.

Serasa. *Educação Financeira e Inadimplência*. 2023. Acesso em: 2024-06-06. Disponível em: <https://www.serasa.com.br/educacao-financeira>.

SILVA, B. C. O. d.; NóBREGA, R. S. Geografia quantitativa, por quê não? *Revista Vozes dos Vales, UFVJM, Diamantina*, v. 14, p. 1–28, 2018. Disponível em: <www.ufvjm.edu.br/vozes>.

SILVA, E.; RIBEIRO, A. P. A aplicação da análise de redes sociais na pesquisa sobre políticas públicas. *Revista Brasileira de Política e Gestão Pública*, v. 6, n. 1, p. 45–59, 2016. Disponível em: <https://www.redalyc.org/articulo.oa?id=>.

SILVA, J. C.; OLIVEIRA, R. P. *Web Scraping em Dados Públicos: Uma Análise Ética e Legal*. 2020. Disponível em: `https://cip.brapci.inf.br/download/170337`. Acesso em: 2024-06-07.

SINKEY, J. F. *Commercial Bank Financial Management*. 6. ed. [S.l.]: Prentice Hall, 2002.

SINKEY, J. F. *Commercial Bank Financial Management in the Financial-Services Industry*. Upper Saddle River: Prentice Hall, 2002.

SMITH, J. *Credit Analysis and Risk Management*. New York: Financial Times Press, 2020.

SOARES, T. C. et al. Pesquisa quantitativa em turismo: os dados gerados são válidos e confiáveis? *Revista Iberoamericana de Turismo- RITUR*, v. 9, n. 1, p. 162–174, 2019.

SPC Brasil. *Como calcular risco de crédito*. 2023. Acesso em: 2023-06-06. Disponível em: <https://www.spcbrasil.org.br/blog/como-calcular-risco-de-credito>.

SPC Brasil. *Como calcular risco de crédito*. 2023. Acesso em: 2023-06-06. Disponível em: <https://www.spcbrasil.org.br/blog/como-calcular-risco-de-credito>.

SPRINGERLINK. A credit scoring model for smes based on social media data. *SpringerLink*, 2024. Disponível em: <https://link.springer.com>.

Superior Tribunal de Justiça. *Documento Jurídico*. 2023. Acesso em: 2023-06-06. Disponível em: <https://processo.stj.jus.br/processo/revista/documento/mediado-/?componente=ITA\&sequencial=2289458&num_registro=202300677939&data=20230427&formato=

TANG, J. et al. Line: Large-scale information network embedding. *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*, ACM, Florence, p. 1067–1077, 2015.

TAYLOR, L. *Conditions and Credit Assessment in the Digital Age*. Oxford: Oxford University Press, 2019.

The Economist. *Economic Impact of Default Rates*. 2023. Acesso em: 2024-06-06. Disponível em: <https://www.economist.com/finance-and-economics/2023/06/06-/economic-impact-of-default-rates>.

THOMAS, L. C.; EDELMAN, D. B.; CROOK, J. N. *Credit Scoring and Its Applications*. Philadelphia: SIAM, 2002.

THOMAS, L. C.; EDELMAN, D. B.; CROOK, J. N. *Credit Scoring and Its Applications*. Philadelphia: SIAM, 2002.

VISTA, B. *Consulta de Crédito e Informações sobre Consumidores*. 2023. Acesso em: 2024-06-07. Disponível em: <https://www.boavistaservicos.com.br/>.

WHARTON, K. at. The surprising ways that social media can be used for credit scoring. *Knowledge at Wharton*, 2024. Disponível em: <https://knowledge.wharton.upenn.edu>.

WHITE, S. *New Approaches in Credit Analysis Using Social Media*. Chicago: University of Chicago Press, 2021.

WILSON, D. *Character and Creditworthiness*. Philadelphia: Wharton School Press, 2015.

YAO, G. et al. Using social media information to predict the credit risk of listed enterprises in the supply chain. *Journal of Financial Services Marketing*, Emerald Publishing Limited, v. 29, n. 2, p. 4993–5016, 2024. Disponível em: <https://www-.emerald.com/insight/content/doi/10.1108/K-12-2021-1376/full/html>.

ÓSKARSDóTTIR, M. et al. The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. *Applied Soft Computing*, v. 74, p. 26–39, 2019.