Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

O uso de técnicas de Aprendizado de Máquina para concessão de crédito: um estudo a partir do portal de dados abertos brasileiro

Luana Thamiris da Silva de Oliveira

Dissertação apresentada como requisito de conclusão do
Mestrado Profissional em Computação Aplicada

Orientador
Prof. Dr. Marcio de Carvalho Victorino

Brasília
2024

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

dO48u

de Oliveira, Luana Thamiris da Silva
O uso de técnicas de Aprendizado de Máquina para concessão de crédito: um estudo a partir do portal de dados abertos brasileiro / Luana Thamiris da Silva de Oliveira; orientador Marcio de Carvalho Victorino. -- Brasília, 2024.
78 p.

Tese(Mestrado Profissional em Computação Aplicada) -- Universidade de Brasília, 2024.

1. Auxílio Emergencial. 2. Modelagem de Crédito. 3. Machine Learning. 4. Instituições Financeiras. 5. Avaliação de Crédito. I. Victorino, Marcio de Carvalho, orient. II. Título.

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

O uso de técnicas de Aprendizado de Máquina para concessão de crédito: um estudo a partir do portal de dados abertos brasileiro

Luana Thamiris da Silva de Oliveira

Dissertação para requisito de conclusão do
Mestrado Profissional em Computação Aplicada

Prof. Dr. Marcio de Carvalho Victorino (Orientador)
CIC/UnB

Prof. Dr. Wallace Anacleto Pinheiro
Exército Brasileiro - EB

Prof. Dr. Andrei Lima Queiroz
STI/UnB

Prof. Dr. Marcelo Ladeira - Suplente
CIC/UnB

Prof. Dr. Gladston Luiz da Silva
Coordenador do Programa de Pós-graduação em Computação Aplicada

Brasília, 28 de setembro de 2024

"All modelos are wrong, but some are useful" - George E.P. Box

Agradecimentos

Gostaria de expressar meus sinceros agradecimentos ao meu orientador, Prof. Dr. Márcio de Carvalho Victorino, cuja orientação, apoio e incentivo foram essenciais ao longo de todo o processo de elaboração deste trabalho. Sua disponibilidade e comprometimento foram fundamentais para que eu pudesse superar os desafios encontrados nesta jornada.

Agradeço também aos membros da banca, Prof. Dr. Marcelo Ladeira e Prof. Dr. Andrei Lima Queiroz, por aceitarem o desafio de avaliar esta pesquisa e pelas valiosas contribuições que enriqueceram meu trabalho.

Um agradecimento especial à instituição onde trabalho e aos meus colegas, que acreditaram na viabilidade deste estudo e na possibilidade de alcançarmos resultados positivos.

Não posso deixar de mencionar a resiliência que todos nós demonstramos ao enfrentar a pandemia de COVID-19. Essa experiência nos ensinou a valorizar ainda mais os momentos de aprendizado e convivência, mostrando que, mesmo em tempos difíceis, é possível encontrar formas de seguir em frente.

Por fim, sou profundamente grata aos meus familiares e amigos, em especial ao meu Sol e Ducanilda, que sempre estiveram ao meu lado, oferecendo amor e suporte incondicional. Sem vocês, este trabalho não teria sido possível. Obrigada por acreditarem em mim e por me inspirarem a dar o meu melhor.

Resumo

Este estudo tem como objetivo principal incorporar os dados do Auxílio Emergencial, concedido pelo Governo Federal do Brasil entre abril e dezembro de 2020, ao modelo de aprendizado de máquina para concessão de empréstimo pessoal adotado por uma instituição financeira brasileira, cujo nome será mantido em anonimato. Essa instituição atende majoritariamente indivíduos das classes C, D e E, que frequentemente enfrentam obstáculos para obter crédito em bancos tradicionais. O estudo adota técnicas avançadas de avaliação de crédito, além das principais métricas de validação de modelos de crédito, visando aprimorar a precisão e a eficácia do modelo em questão. Inicialmente, os modelos foram gerados com base nos dados internos da instituição, sendo, posteriormente, incluídos os dados do Auxílio Emergencial para avaliar seu impacto na acurácia, na métrica KS e em outras métricas relevantes. A fundamentação teórica, incluindo técnicas de Regressão Logística, Florestas Aleatórias, LightGBM, modelos com custos e métodos para correção de desbalanceamento, foi embasada em revisão bibliográfica. O ciclo de projeto CRISP-DM orientou a implementação do estudo de caso, cujos resultados demonstraram uma melhora significativa no desempenho do modelo após a inclusão dos dados do auxílio, com aumento de 5,6 pontos percentuais na métrica KS e de 13,2 pontos percentuais na acurácia.

Palavras-chave: Aprendizado de máquina, avaliação de crédito, modelagem de crédito, auxílio emergencial, inclusão financeira, instituições financeiras

Abstract

This study aims to incorporate data from the Emergency Aid provided by the Brazilian Federal Government between April and December 2020 into the machine learning model for personal loan approval used by an anonymous Brazilian financial institution. This institution primarily serves individuals in socio-economic classes C, D, and E, who often face obstacles in obtaining credit from traditional banks. The study employs advanced credit evaluation techniques and leading credit model validation metrics to enhance the model's accuracy and effectiveness. Initially, models were developed using the institution's internal data, with Emergency Aid data later incorporated to assess its impact on accuracy, KS metric, and other relevant metrics. The theoretical foundation, including Logistic Regression, Random Forest, LightGBM, cost-sensitive models, and methods for addressing imbalance, was based on a literature review. The CRISP-DM project cycle guided the implementation of the case study, with results showing a significant improvement in model performance after incorporating the aid data, with an increase of 5.6 percentage points in the KS metric and 13.2 percentage points in accuracy.

Keywords: Machine learning, credit evaluation, credit modeling, emergency aid, financial inclusion, financial institutions

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Definição do Problema

Em 2020, cerca de 21,9 milhões de pessoas no Brasil, representando 10,4% da população, viviam com uma renda de até ¼ do salário mínimo per capita mensal, o equivalente a aproximadamente R$ 261,00. Além disso, 29,1% da população, ou cerca de 61,4 milhões de pessoas, tinham uma renda per capita de até ½ salário mínimo, aproximadamente R$ 522,00 [5].

Com uma parte significativa da população vivendo com renda mínima de R$ 522,00, agravada pela crise econômica decorrente da pandemia da Covid-19, o governo federal instituiu o Auxílio Emergencial. De acordo com o Decreto Nº 10.316, de 7 de abril de 2020, o Auxílio Emergencial, aprovado pelo Congresso Nacional e sancionado pelo Presidente da República, foi inicialmente concedido por 3 meses, com o valor de R$ 600,00 para pessoas que se enquadrassem em determinados critérios [6]. Após esse período inicial, o auxílio foi reduzido para R$ 350,00, estendendo-se até o final de 2020.

Geralmente, esta parcela da população, que teve acesso ao Auxílio Emergencial, enfrenta dificuldades para acessar ofertas de crédito das instituições financeiras. Isso ocorre porque, frequentemente, essas pessoas não conseguem comprovar renda formal, como no caso dos trabalhadores autônomos, ou possuem rendimentos abaixo do mínimo exigido por tais instituições. Além disso, as características financeiras e sociais homogêneas desse grupo resultam em uma classificação de crédito que desqualifica todos para o acesso ao crédito.

Diante desse cenário, o presente estudo objetiva incorporar os dados do Auxílio Emergencial na análise das variáveis que compõem um modelo de aprendizado de máquina para concessão de empréstimo pessoal em um banco digital brasileiro.

As instituições financeiras, incluindo bancos, desempenham o papel de captar recursos e conceder crédito, além de oferecer outros serviços financeiros como saques, empréstimos

e investimentos. No Brasil, os bancos são regulados pelo Banco Central (BC), que supervisiona o cumprimento das normas que regem o Sistema Financeiro Nacional (SFN) [7].

O SFN é formado por instituições e entidades que facilitam a intermediação financeira, conectando credores e tomadores de recursos. O sistema inclui agentes normativos, supervisores e operadores, sendo os operadores responsáveis pela oferta de serviços financeiros.

Os bancos digitais, que incluem as Sociedades de Crédito Direto (SCD) e as Sociedades de Empréstimo entre Pessoas (SEP), são exemplos de operadores dentro do SFN, regulamentados pela Resolução 4.656, de 26 de abril de 2018. Essas entidades operam exclusivamente de forma digital e baseadas em tecnologia [8].

As SCDs realizam operações de empréstimo, financiamento e aquisição de direitos creditórios utilizando recursos próprios, exclusivamente por meio de plataformas eletrônicas. As SEPs, conhecidas como empréstimo de pessoa para pessoa (P2P), conectam diretamente investidores e tomadores de crédito através de plataformas seguras do ponto de vista tecnológico, jurídico e mercadológico [8].

Anteriormente à regulamentação que permitiu sua incorporação no SFN, muitos bancos digitais atuavam como correspondentes de instituições financeiras tradicionais, mas já apresentavam inovações em termos de tecnologia (e.g., plataformas online, inteligência artificial, grandes bases de dados) e relacionamento com clientes exclusivamente por canais eletrônicos, sem necessidade de presença física [8].

Com o crescimento dos bancos digitais, surge a necessidade de mitigar riscos financeiros, especialmente o risco de crédito, definido como a probabilidade de inadimplência por parte do tomador. Para identificar potenciais riscos, as instituições utilizam diversas informações do cliente, como idade, renda e tempo de conta, que resultam na pontuação de crédito, essencial para as decisões de concessão de crédito. Modelos de aprendizado de máquina têm se mostrado ferramentas valiosas nesse contexto devido à sua versatilidade e precisão.

Os modelos baseados em aprendizado de máquina correlacionam características individuais do mutuário com o risco de inadimplência, permitindo classificar clientes em bons ou maus pagadores e, assim, apoiar as decisões financeiras [9].

Entretanto, muitos desses modelos falham em discriminar corretamente os beneficiários do Auxílio Emergencial, agrupando-os em uma categoria de pontuação de crédito que frequentemente resulta na recusa de empréstimos.

Ao integrar dados do Auxílio Emergencial nos modelos de concessão de crédito, espera-se aumentar a precisão das análises, possibilitando atender um público com necessidades creditícias que não são supridas pelas instituições tradicionais.

1.2 Justificativa

A proposta deste trabalho é utilizar dados do modelo de concessão de crédito adotado por uma instituição financeira brasileira, que por motivos de privacidade dos dados, não terá o nome revelado. Fundada em meados de 2019, a instituição tem a missão de proporcionar serviços financeiros a quem realmente necessita, especialmente aqueles que não têm acesso aos bancos tradicionais. Para alcançar esse objetivo, ela oferece serviços financeiros acessíveis, eficientes e gratuitos à população brasileira.

O ciclo de crédito geralmente é composto por quatro etapas, cada uma delas representada por um modelo de aprendizado de máquina. A primeira etapa envolve o modelo de prospecção (*application score*), que é utilizado pelas instituições financeiras para identificar o perfil de clientes mais aderente ao produto ofertado. A segunda etapa, que é o foco deste trabalho, está relacionada à concessão de crédito para esses novos clientes, utilizando modelos de crédito (*credit score*) para avaliar a probabilidade de inadimplência. A terceira fase refere-se ao gerenciamento da carteira de clientes, onde são aplicados modelos de comportamento (*behavior score*) que avaliam o histórico dos clientes. Finalmente, a última etapa envolve modelos de cobrança (*collection score*), que estimam a probabilidade de recuperação da dívida entre clientes inadimplentes [1].

Atualmente, o modelo de concessão de empréstimo pessoal da instituição financeira, que é o objeto deste estudo, é composto por 16 variáveis explicativas internas e uma variável resposta. Esse modelo foi desenvolvido com dados de todos os clientes que realizaram cadastro no aplicativo da instituição entre junho e dezembro de 2020, período marcado pelo auge da pandemia de Covid-19.

Com o tempo, como era previsível, o modelo sofreu um processo de perda de performance preditiva, perdendo parte da sua capacidade de segmentar o público com a eficiência previamente atingida. Considerando a proposta da instituição financeira de atender uma parcela específica da população brasileira e o fato de que essa parcela é mais sensível e impactada por variações no cenário macroeconômico, identificou-se uma oportunidade para explorar a aplicação de variáveis externas de grande impacto social, como os dados do Auxílio Emergencial. A inclusão desses dados tem o potencial de recalibrar o modelo e aumentar sua precisão na concessão de crédito, especialmente para esses segmentos populacionais mais vulneráveis.

1.3 Perguntas da Pesquisa

Como aprimorar um modelo de concessão de crédito para um público na maioria das vezes homogêneo entre si, com o intuito de determinar os bons e os maus pagadores?

1.4 Hipótese da Pesquisa

A utilização de informações governamentais, como os dados do Auxílio Emergencial, tem o potencial de melhorar significativamente o desempenho do modelo de concessão de crédito para um público que frequentemente enfrenta dificuldades para comprovar renda ou possui uma renda inferior à média nacional. A inclusão desses dados pode permitir uma segmentação mais precisa, contribuindo para o aprimoramento das políticas de crédito adotadas pela instituição financeira.

Ao combinar as variáveis internas com as informações do Auxílio Emergencial, espera-se desenvolver um modelo que considere de forma mais abrangente as características de indivíduos pertencentes às classes C, D e E, levando em conta suas especificidades que são comumente negligenciadas pelos modelos de crédito tradicionais. Essa abordagem visa ajustar as práticas de concessão de crédito para atender melhor a um segmento da população que historicamente tem suas necessidades creditícias sub-atendidas, promovendo assim uma inclusão financeira mais justa e eficaz.

1.5 Objetivo Geral

Este trabalho tem por objetivo propor e avaliar a incorporação dos dados do Auxílio Emergencial no estudo das variáveis que compõem o modelo de aprendizado de máquina para concessão de empréstimo pessoal para os públicos da classe C, D e E. Esses dados possibilitarão uma discriminação maior do público, permitindo a adoção de políticas de crédito mais assertivas e minimizando o risco de crédito da instituição.

1.6 Objetivos Específicos

Para alcançar o objetivo geral proposto, os seguintes objetivos específicos foram estabelecidos:

- Analisar diferentes técnicas de aprendizado de máquina e compará-las em termos das métricas para adequação de ajuste de um modelo de concessão de crédito;
- Analisar e validar a base de dados fornecida no Portal de Dados Abertos do Governo Federal com os dados do Auxílio Emergencial no ano de 2020;

1.7 Contribuição da Pesquisa

Entende-se que a partir da combinação das variáveis internas mais as informações do Auxílio Emergencial, será possível elaborar um modelo considerando as características

das pessoas das classes C, D e E, de forma à considerar suas especificidades normalmente desconsideradas nos modelos de crédito tradicionais.

1.8 Estrutura do Trabalho

O presente trabalho está estruturado em outros 5 capítulos, além da presente Introdução:

- O Capítulo 2 apresenta a metodologia da pesquisa, composta por uma revisão da literatura contendo os principais tópicos relacionados à fundamentação teórica;
- O Capítulo 3 aborda trabalhos relacionados a esta pesquisa, apontando os pontos positivos, as dificuldades e limitações nesse tipo de estudo e seus objetivos;
- O Capítulo 4 traz informações sobre a fundamentação teórica, com os principais conceitos utilizados para o desenvolvimento deste trabalho. São abordados os conceitos relacionados as técnicas de aprendizado de máquina e os principais modelos adotados nas instituições financeiras. Além disso, o capítulo retrata como funciona o ciclo de crédito, o que são os dados abertos e como o uso dessas informações pode agregar nas análises nos modelos de aprendizado de máquina para concessão de crédito;
- O Capítulo 5 apresenta o estudo de caso, com a descrição das etapas iniciais do CRISP-DM, entendimento do negócio, entendimento dos dados, preparação dos dados e resultados da modelagem;
- O Capítulo 6 apresenta as conclusões da pesquisa realizada, assim como, as discussões relacionadas à confirmação da hipótese do trabalho apresentado no Capítulo 1 e verificando o cumprimento dos objetivos gerais e específicos apresentados também no Capítulo 1 deste estudo.

Capítulo 2

Metodologia

Com o intuito de se levantar os principais estudos relacionados à temática deste trabalho, bem como, os principais tópicos referentes à fundamentação teórica, adotou-se a revisão sistemática da literatura.

2.1 Revisão Bibliográfica da Literatura

A revisão da bibliografia deste projeto foi embasado na metodologia da revisão sistemática que na literatura é um método para conduzir a revisão bibliográfica de maneira formal, seguindo etapas bem definidas, permitindo ainda que os resultados sejam reprodutíveis [10]. Ao contrário de uma revisão bibliográfica tradicional, que utiliza uma seleção de literatura *ad hoc*, a revisão sistemática (RS) é uma revisão metodologicamente rigorosa dos resultados da pesquisa. O objetivo de uma RS não é apenas agregar todas as evidências existentes sobre uma questão de pesquisa; ela também se destina a apoiar o desenvolvimento de diretrizes baseadas em evidências [11].

A aplicação da revisão sistemática envolve três principais fases: planejamento, condução e apresentação de resultados. Na primeira fase, é definido um protocolo de revisão, no qual as perguntas de pesquisa são especificadas juntamente com as palavras-chaves. Depois disso, na segunda fase, o protocolo de revisão é aplicado e a informação é extraída das referências devolvidas. Referências utilizadas para a extração de informações são chamadas de estudos primários, enquanto a revisão é um estudo secundário. Por fim, a terceira fase irá definir como apresentar os resultados da pesquisa [10].

No âmbito das instituições financeiras digitais, em 2019, Milian et.al publicaram um artigo com o objetivo de investigar o conceito de bancos digitais, identificar as literaturas relacionadas ao tema e apontar novos caminhos e oportunidades. Para tal, realizaram uma

revisão da literatura tentando descrever as áreas de atividades das instituições financeiras digitais, verificando que esse setor surge como empresas inovadores ativas principalmente na indústria financeira [12].

Neste trabalho, a revisão foi separada em três temas: Ciclo de crédito, Dados abertos e Técnicas de Aprendizado de Máquina. Para seleção dos artigos utilizou-se como fonte científica o *Scopus* e todas as buscas realizadas foram para trabalhos publicados em inglês e a partir de 2014. Destaca-se ainda, que os resultados obtidos para os três temas refletem uma busca realizada em outubro de 2022. As etapas e critérios utilizados nesse processo serão apresentados a seguir.

A primeira etapa desse processo foi determinar para cada um dos temas as questões de pesquisas que os artigos poderiam conter e uma lista de palavras-chaves nos quais a busca poderia ocorrer. Em seguida, foi elaborada a equação de busca no *scopus* e a seguinte regra deveria ser seguida, conforme mostra Figura 2.1:

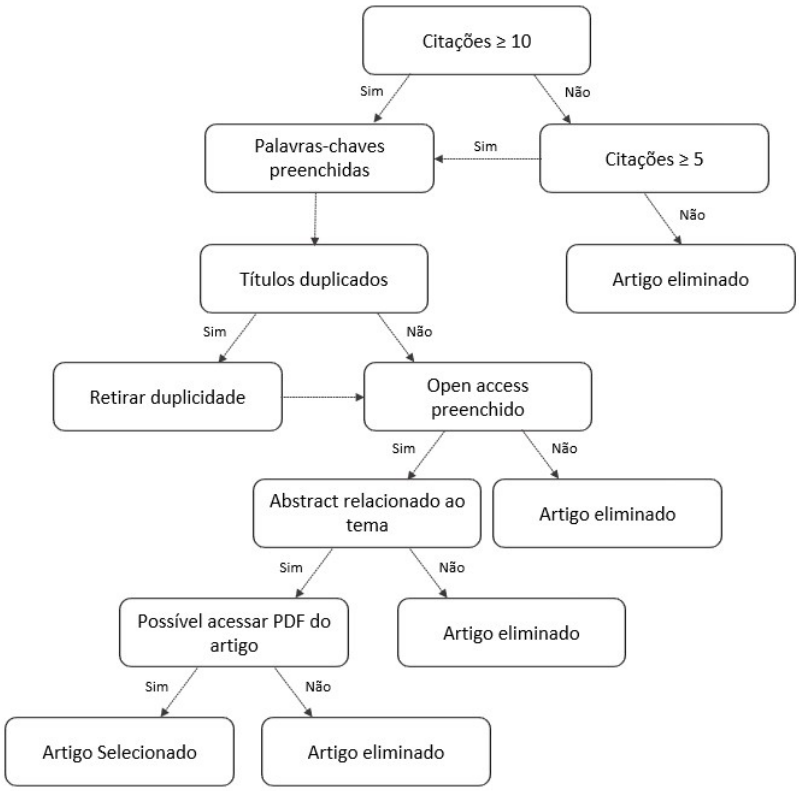


Figura 2.1: Regras para escolha dos artigos no *scopus*.

2.1.1 Ciclo de Crédito

Para o Ciclo de Crédito foram elaboradas cinco perguntas (Tabela 2.1) e uma lista com 9 palavras-chaves que compuseram a busca. Para esse tema foram necessárias três regras que serão explicitadas a seguir.

Tabela 2.1: Questões da pesquisa para Ciclo de Crédito.

Item	**Pergunta**
1	O que é ciclo de crédito nas fintechs?
2	Quais variáveis são utilizadas para análise de crédito?
3	O que é risco de crédito?
4	Quais métricas são utilizadas para avaliar crédito nas fintechs?
5	Como definir o perfil dos clientes que receberão crédito?

Das perguntas levantadas na Tabela 2.1 foram listadas nove palavras-chaves:

1. Ciclo de crédito
2. Crédito
3. Análise de crédito
4. Análise de risco
5. Risco
6. Risco de crédito
7. Score de crédito
8. Avaliar crédito
9. Perfil de clientes de crédito

De posse das perguntas e das palavras-chaves, criou-se a primeira sintaxe (Equação 2.1) de busca cujo retorno compreendeu 1,695.195 artigos, inviabilizando a análise.

$$\text{(TITLE-ABS-KEY (credit AND score) OR TITLE-ABS-KEY (credit AND analysis) OR TITLE-ABS-KEY (cycle AND of AND credit) OR TITLE-ABS-KEY (credit) OR TITLE-ABS-KEY (bank AND score) OR TITLE-ABS-KEY (bank AND credit) OR TITLE-ABS-KEY (score))} \tag{2.1}$$

Foi necessário então, refinar um pouco mais a busca. Nesse momento, foram adicionados os critérios de artigos publicados em inglês e com data a partir de 2014. A Equação 2.2 teve como retorno 872.506 artigos publicados.

$$\begin{aligned}&(\text{TITLE-ABS-KEY (credit AND score) OR TITLE-ABS-KEY (credit AND analysis) OR}\\&\text{TITLE-ABS-KEY (cycle AND of AND credit) OR TITLE-ABS-KEY (credit) OR}\\&\text{TITLE-ABS-KEY (bank AND score) OR TITLE-ABS-KEY (bank AND credit) OR}\\&\text{TITLE-ABS-KEY (score) AND LANGUAGE (english)) AND PUBYEAR} > 2014\end{aligned} \tag{2.2}$$

Tal volumetria foi considerada ainda como inviável para análise das etapas mencionadas anteriormente na Figura 2.1. Tendo em vista que o escopo dos dados deste trabalho referem-se a uma instituições financeiras digitais, acrescentou-se na busca esta palavra. A Equação 2.3 a seguir mostra a sintaxe final cujo resultado foram 385 artigos relacionados ao tema.

$$\begin{aligned}&(\text{TITLE-ABS-KEY (credit AND score) OR TITLE-ABS-KEY (credit AND analysis) OR}\\&\text{TITLE-ABS-KEY (cycle AND of AND credit) OR TITLE-ABS-KEY (credit) OR}\\&\text{TITLE-ABS-KEY (bank AND score) OR TITLE-ABS-KEY (bank AND credit) OR}\\&\text{TITLE-ABS-KEY (score) AND LANGUAGE (english) AND TITLE-ABS-KEY (fintech)) AND}\\&\text{PUBYEAR} > 2014\end{aligned} \tag{2.3}$$

Os 385 artigos foram extraídos do *scopus* com as seguintes informações: autor(es), título, ano de publicação, título de busca, número de citações, *hiperlink* do artigo, resumo do artigo, palavras-chaves e *open access*. A Figura 2.2 a seguir traz uma informação gerada no *scopus* que mostra a evolução de artigos relacionados ao tema apresentados ao longo dos anos.

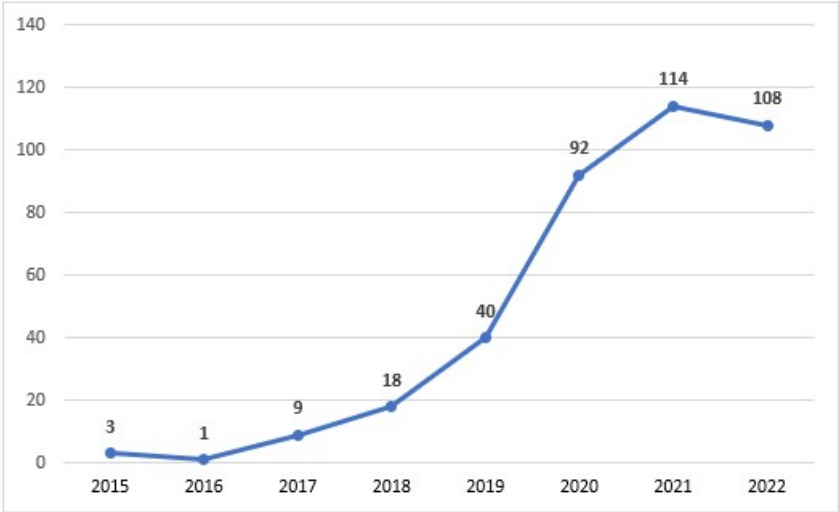


Figura 2.2: Evolução de artigos publicados sobre ciclo de crédito *scopus*.

A etapa seguinte consistiu em aplicar as regras apresentadas na Figura 2.1 e eliminar os artigos que não atendiam aos critérios previamente estabelecidos. A Tabela 2.2, resume a quantidade de artigos em cada uma dessas etapas até a seleção final.

Tabela 2.2: Artigos que atendem aos critérios estabelecidos - Ciclo de Crédito.

Regras pré-estabelecidas	**Quantidade de artigos**
Citações >= 10	57
Palavras-chaves preenchidas	56
Títulos duplicados	55
Open Access preenchido	30
Abstract diretamente relacionado ao tema	27
Acesso ao *PDF* finalizado	22

2.1.2 Dados Abertos

O segundo tema foi sobre dados abertos. Aplicando-se a mesma lógica mencionada, foram elaboradas perguntas e uma lista de palavras-chaves, apresentadas a seguir.

A Tabela 2.3 apresenta quatro perguntas relacionadas aos dados abertos. São questões que visam responder quais informações estão disponíveis para consultas e em quais fontes podem ser utilizadas.

Tabela 2.3: Questões da pesquisa para Dados Abertos.

Item	**Pergunta**
1	O que são dados abertos?
2	Quais dados estão disponíveis para consultas?
3	Quais fontes podemos buscar dados abertos?
4	Quais dados são disponíveis pelo governo?

A partir das perguntas apresentadas, listou-se as palavras-chaves que comporiam a busca no *scopus* e que auxiliariam no desenvolvimento dessa dissertação:

1. Dados abertos
2. Fontes de dados
3. Dados do governo
4. Portal da transparência
5. Acesso a dados públicos

A primeira sintaxe (Equação 2.4) elaborada para este tema é apresentada a seguir:

$$\text{TITLE-ABS-KEY (dados AND abertos) OR TITLE-ABS-KEY (fontes AND de AND dados AND abertos) OR TITLE-ABS-KEY (dados AND do AND governo) OR TITLE-ABS-KEY (portal AND da AND transparência) OR TITLE-ABS-KEY (acesso AND à AND dados AND públicos) AND LANGUAGE (english) AND PUBYEAR} > 2014 \tag{2.4}$$

No entanto, o total de artigos retornados a partir desta busca foi 1, o que significa que a Equação 2.4 foi muito restrita para o tema. Na tentativa de abranger um quantitativo maior de artigos sobre dados abertos retirou-se da sintaxe os critérios de artigos com publicação em língua inglesa e que fossem a partir de 2014. A Equação 2.5 foi então aplicada no *scopus* e foram encontrados 103 documentos no total.

TITLE-ABS-KEY (dados AND abertos) OR TITLE-ABS-KEY (fontes AND de AND dados AND abertos) OR TITLE-ABS-KEY (dados AND do AND governo) OR TITLE-ABS-KEY (portal AND da AND transparência) OR TITLE-ABS-KEY (acesso AND à AND dados AND públicos)

(2.5)

Para elucidar a evolução, bem como a dificuldade de se encontrar artigos publicados com este tema, extraiu-se do *scopus* a quantidade de artigos publicados a partir de 2001 até 2022. Ressalta-se que esses números refletem a quantidade de trabalhos publicados até a data de realização das consultas, podendo sofrer alguma alteração em 2022 para consultas posteriores a outubro. A Figura 2.3 detalha essa evolução ao longo do tempo:

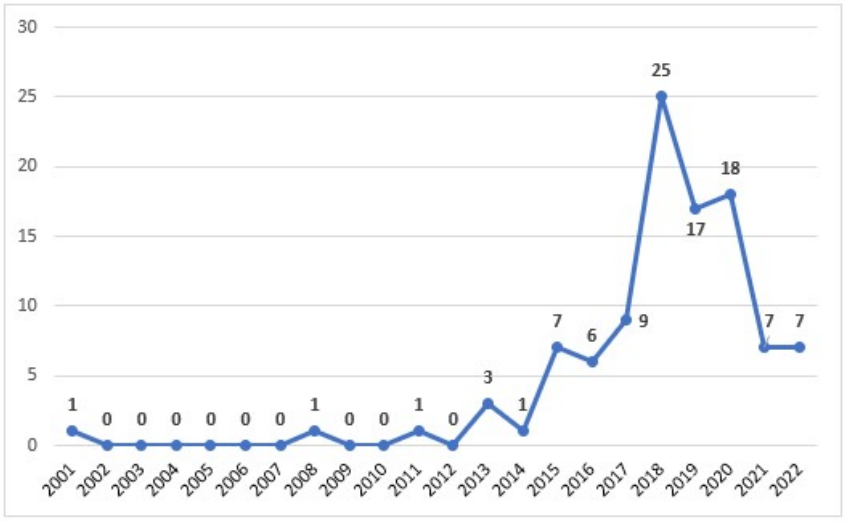


Figura 2.3: Evolução de artigos publicados sobre dados abertos *scopus*.

Os 103 artigos foram extraídos do *scopus* também com as informações de autor(es), título, ano de publicação, título de busca, número de citações, *hiperlink* do artigo, resumo do artigo, palavras-chaves e *open access* para que fossem aplicadas as regras mencionadas na Figura 2.1. Os resultados da aplicação da regra foram sumarizados na Tabela 2.4 a seguir:

Tabela 2.4: Artigos que atendem aos critérios estabelecidos - Dados Abertos.

Regras pré-estabelecidas	**Quantidade de artigos**
Citações >= 5	19
Palavras-chaves preenchidas	12
Títulos duplicados	12
Open Access preenchido	9
Abstract diretamente relacionado ao tema	5
Acesso ao *PDF* finalizado	4

2.1.3 Técnicas de Aprendizado de Máquina

O último tema proposto para aplicar a revisão da literatura foi sobre as Técnicas de Aprendizado de Máquina. Assim como para os dois temas anteriores, foram elaboradas questões de pesquisa e uma lista de palavras-chaves. Destaca-se, que para este tópico na consulta do *Scopus* foi utilizado *Machine Learning* como expressão de consulta no lugar da expressão em português, Aprendizado de Máquina, traduzida na Tabela 2.5 a seguir:

Tabela 2.5: Questões da pesquisa para Técnicas de Aprendizado de Máquina.

Item	**Pergunta**
1	Quais as técnicas de *Aprendizado de Máquina* mais utilizadas na área de crédito?
2	O que são técnicas de *Aprendizado de Máquina*?
3	Como os bancos e *fintechs* usam técnicas de *Aprendizado de Máquina*?
4	Quais métricas existem para escolher a melhor técnica?
5	Quais técnicas são utilizadas na modelagem de crédito?

A partir das perguntas apresentadas, listou-se as palavras-chaves que comporiam a busca no *scopus*:

1. Técnicas de *Machine Learning*
2. Algoritmos
3. Algoritmos de *Machine Learning*
4. *Machine Learning*
5. Modelagem de crédito
6. Modelos de crédito
7. Técnicas de *Machine Learning* em bancos
8. Técnicas de *Machine Learning* nas *fintechs*

9. Métricas de avaliação de técnicas de *Machine Learning*

Para este tema apenas uma consulta foi necessária, retornando um total de 405 artigos a serem analisados nas regras pré-estabelecidas. A Equação 2.6 a seguir mostra a sintaxe utilizada na consulta de artigos sobre técnicas de *Aprendizado de Máquina*:

TITLE-ABS-KEY (modelagem AND de AND crédito) OR TITLE-ABS-KEY (algoritmos) OR TITLE-ABS-KEY (modelos AND de AND machine AND learning) OR TITLE-ABS-KEY (modelagem AND de AND crédito AND nas AND fintechs) OR TITLE-ABS-KEY (técnicas AND de AND machine AND learning AND nos AND bancos) OR TITLE-ABS-KEY (técnicas AND de AND machine AND learning AND nas AND fintechs) OR TITLE-ABS-KEY (métricas AND de AND avaliação AND de AND modelos) OR TITLE-ABS-KEY (algoritmos AND de AND machine AND learning) OR TITLE-ABS-KEY (técnicas AND de AND machine AND learning) AND LANGUAGE (english) AND PUBYEAR > 2014 (2.6)

Os 405 artigos foram extraídos do *scopus* com as informações de autor(es), título, ano de publicação, título de busca, número de citações, *hiperlink* do artigo, resumo do artigo, palavras-chaves e *open access* para que fossem aplicadas as regras mencionadas na Figura 2.1. Os resultados da aplicação da regra foram sumarizados na Tabela 2.6 a seguir:

Tabela 2.6: Artigos que atendem aos critérios estabelecidos - Técnicas de *Aprendizado de Máquina*.

Regras pré-estabelecidas	**Quantidade de artigos**
Citações >= 5	65
Palavras-chaves preenchidas	64
Títulos duplicados	63
Open Access preenchido	40
Abstract diretamente relacionado ao tema	15
Acesso ao *PDF* finalizado	15

A Figura 2.4 a seguir mostra a evolução de artigos publicados

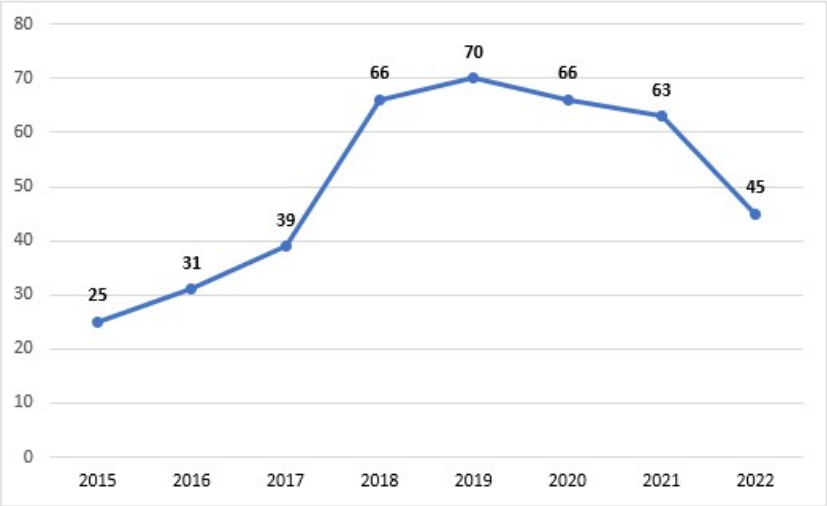


Figura 2.4: Evolução de artigos publicados sobre Técnicas de *Aprendizado de Máquina* *scopus*.

Conclui-se então, que para a temática de crédito 22 artigos serão tidos como referência e apresentados no capítulo de fundamentação teórica. Para os temas de dados abertos e técnicas de aprendizado de máquina serão 4 e 15 artigos, respectivamente.

Capítulo 3

Trabalhos Relacionados

Neste capítulo, serão apresentados alguns trabalhos relacionados à pesquisa. Serão apresentados artigos sobre o uso de dados abertos, ciclo de crédito e técnicas de aprendizado de máquina aplicadas ao crédito.

Após a realização da revisão da literatura foram totalizados 41 artigos, distribuídos da seguinte maneira: 22 relacionados ao ciclo de crédito, 4 sobre dados abertos e 15 artigos sobre técnicas de aprendizado de máquina aplicadas ao crédito.

3.1 Ciclo de Crédito

Sicsú [3], destaca que a concessão de crédito é uma decisão sob condições de incerteza. O risco de uma solicitação de crédito pode ser avaliado de forma subjetiva ou medido de forma quantitativa. Destaca ainda que a avaliação subjetiva, apesar de incorporar a experiência do analista, não quantifica o risco de crédito. Por sua vez, utilizar técnicas quantitativas apresenta uma série de vantagens como: consistência, decisões rápidas e mais adequadas, bem como, a possibilidade de monitorar e administrar o risco de um portfólio de crédito.

Forti [1] ressalta ainda que, o risco de crédito pode ser definido como a probabilidade de perda de um empréstimo financeiro e, por isso, as empresas utilizam métodos subjetivos e/ou quantitativos para obter uma decisão mais confiável. Devido a esse fato, surge o ciclo de crédito, o qual busca medir o risco dos clientes em diferentes fases de relacionamento com a instituição.

Niklas, Giudici et al. [13] afirmam que o risco de crédito pode ser medido com modelos de aprendizado de máquina muito diferentes, capazes de extrair relações não lineares

entre as informações financeiras dos balanços. Os autores propuseram um modelo de inteligência artificial a ser aplicado no gerenciamento de riscos dos bancos digitais no qual revelaram que tanto os mutuários arriscados quanto os não arriscados podem ser agrupados de acordo com um conjunto de características financeiras similares, que podem ser empregadas para explicar e compreender sua pontuação de crédito e, portanto, para prever seu comportamento futuro, perfazendo assim o ciclo de crédito.

Leong, Tan et al. [14], definiram as instituições financeiras digitais como uma tecnologia financeira que envolve o projeto e a entrega de produtos e serviços financeiros por meio da tecnologia. Tal modelo de negócio, afetou as instituições financeiras tradicionais, reguladores, clientes e comerciantes em uma ampla gama de indústrias. As tecnologias digitais desafiaram os fundamentos do setor financeiro altamente regulamentado, levando ao surgimento de sistemas de pagamento não tradicionais, trocas de dinheiro entre pares e turbulência crescente nos mercados de moedas.

Giudici, Branka et al. [15] e Gruin, Knaack também destacaram que a intermediação financeira mudou muito nas últimas décadas. No contexto dos serviços de crédito, os bancos digitais introduziram muitas oportunidades, entre as quais a melhoria da velocidade, melhor experiência do cliente e custos reduzidos, impactando em todo o ciclo de crédito vigente. Najib, Ermawati et al. [16], Appiah-Otoo, Song [17] e Leong, Tan et al. [14], salientam que, esses bancos se desenvolveram rapidamente em contextos distintos, oferecendo novos produtos e serviços por meio da tecnologia digital, o que permitiu, por exemplo, reduzir a lacuna de financiamento para pequenas empresas, introduzindo novos modelos de negócios, provocando ainda uma melhora nos serviços das instituições financeiras existentes, permitindo a inclusão financeira de segmentos de mercado anteriormente excluídos.

A título de contextualização, os empréstimos pessoa para pessoa (*Peer-to-Peer* - P2P) pertencem aos serviços das instituições digitais que combinam diretamente os emprestadores com os mutuários através de plataformas onlines sem a intermediação de instituições financeiras, tais como os bancos [18]. O surgimento dos empréstimos onlines P2P não só mudou o negócio de empréstimos, mas também trouxe novos tipos de riscos [19]. Ahelegbey, Giudici et al. [20] em seu estudo destacaram que estas melhorias e avanços podem vir ao preço de uma pior medição do risco de crédito, o que pode prejudicar os credores e colocar em risco a estabilidade de um sistema financeiro.

Para mitigar tal risco, Ahelegbey, Giudici et al., propuseram um modelo de pontuação baseado em rede, inferindo nas conexões entre os participantes P2P, baseado na descoberta

de estruturas comunitárias dentro deles. Em complemento, Bernards [21] salienta que o crescimento das instituições digitais não necessariamente significa um acesso maior ao crédito, pois muitas dessas tecnologias inovadoras não levam em conta o caráter desigual e altamente limitado da "inclusão financeira"na prática. Diante disso, Demirgüç-Kunt, Klapper et al [22], a partir da análise de dados da Global Findex, destacaram que é importante medir e acompanhar a inclusão financeira, não apenas para entendê-la, mas também para identificar oportunidades de expansão.

3.2 Dados Abertos

Alcantara, Bandeira et al.[23] destacaram que a publicação de dados abertos tem gerado benefícios em diversas áreas, como a transparência em órgãos governamentais, utilizada como uma importante ferramenta no combate a corrupção e a ampliação da participação da sociedade no processo de desenvolvimento de novas soluções. Além do setor governamental, a educação também torna-se uma beneficiária, pois a grande quantidade de dados disponíveis pode auxiliar na tomada de decisão de professores e gestores escolares. Salientam ainda que ao aplicar dados abertos conectados em contextos educacionais, é possível melhorar a qualidade de conteúdos publicados, melhorar a veracidade das informações, e agilizar o processo de recuperação de conteúdo educacional.

Pavão, Rocha et al.[24] exploraram o cenário nacional e internacional com intuito de encontrar uma solução tecnológica para efetivar o acesso aberto a dados de pesquisa (AADP) dentro de uma perspectiva nacional, uma vez que estes possuem grande importância, não só para validar os resultados obtidos e publicados, mas pela possibilidade do reuso, como também para impulsionar novas pesquisas e socializar o conhecimento. Destacam ainda que, os repositórios de dados de pesquisa assumem o papel de fornecer mecanismos de busca eficientes e serviços de valor agregado para a produção de conhecimento.

Em uma outra frente no uso dos dados abertos, Santos Neto, Marconde et al.[25], propuseram um caso fictício, a partir das obras de Machado de Assis, para a interligação de dados provenientes de arquivos, bibliotecas e museus. O objetivo foi, por meio da identificação de vocabulários já existentes, ampliar a semântica dos conteúdos publicados e da descrição dos conteúdos em Estruturas de Descrição de Recursos (*Resource Description Framework* - RDF) , mostrando, dessa forma, que a interligação dos dados é possível e útil.

Neves, Machado et al. [26] publicaram um estudo sobre os desdobramentos da pandemia de Covid-19 em relação ao desemprego, pobreza e a fome no Brasil. Para tal, utilizaram dados sobre a taxa de desocupação, solicitações de seguro-desemprego e contingente de famílias em extrema pobreza e em insegurança alimentar, coletados em sistemas de informação governamentais, em pesquisas publicadas por órgãos públicos, em artigos científicos e em portais de notícias.

3.3 Técnicas de Aprendizado de Máquina

Feng, Xiao et al. [27], destacam que a ideia principal da pontuação de crédito é construir um modelo quantitativo para estimar o valor do crédito de um cliente, baseado em um conjunto de variáveis explicativas. Durante as últimas décadas, vários algoritmos de classificação têm sido explorados para a pontuação de crédito com base em técnicas estatísticas tradicionais ou técnicas de aprendizado de máquina.

Soui, Gasmi et al. [28], dividiram as técnicas de aprendizado de máquina em duas categorias: i) métodos tradicionais e ii) métodos baseados em regras. A primeira categoria inclui métodos como regressão logística, rede neural, máquinas de vetores de suporte (*support vector machine* - SVM), dentre outros. A segunda categoria combinou diferentes métodos para propor modelos de avaliação baseados em regras. O desempenho dessas várias técnicas é determinado pela precisão dos modelos extraídos.

García, Marqués et al. [29], salientam que, ao contrário dos modelos estatísticos, a aprendizagem de máquinas e os métodos de inteligência computacional não assumem nenhum conhecimento prévio específico, mas, em vez disso, extraem automaticamente informações de observações passadas. No respectivo estudo, analisaram informações de 14 bancos de dados financeiros com o objetivo de comparar o desbalanceamento das classes positivas e negativas com o desempenho de classificadores de agregação (*Bagging*), aprimoramento adaptativo (*AdaBoost*), florestas aleatórias (*Random Forest*), aprimoramento gradiente estocástico (*Stochastic Gradient Boosting*) dentre outros. Os resultados mostraram que, o desempenho dos modelos dependente predominantemente das amostras positivas.

Abellán e Castellano [30] incorporaram ao estudo de García, Marqués et al. [29] a técnica de árvore de decisão credal (*Credal Decision Tree* - CDT) no qual identificaram que o classificador simples pode ser uma escolha interessante para ser usada em base de conjuntos para pontuação de crédito e previsão de falência, provando que não apenas

o desempenho individual de um classificador é o ponto-chave a ser selecionado. Em outro estudo, Abellán e Mantas [31] mostraram que os resultados via CDT de fato foram melhores do que os procedimentos clássicos para dados relacionados à previsão de falência e à pontuação de crédito.

Xia, Zhao et al. [32], propuseram um modelo de crédito baseado em árvores denominado como um comitê heterogêneo cauteloso de superajuste (*Overfiting-cautious heterogeneous ensemble* - OCHE), que envolve cinco algoritmos também baseados em árvores, a saber, florestas aleatórias (*Random Forest* - RF), árvore de decisão por aprimoramento gradiente (*Gradient Boostind Decision Tree* - GBDT), aprimoramento gradiente extremo (*Extreme Gradient Boosting* - XGBoost), modelo de aprimoramento de gradiente leve (textitLight Gradient Boosting Model - LightGBM) e aprimoramento de variáveis categóricas (*Categorical Boosting* - CatBoost). Os resultados mostraram que o modelo proposto pelos autores foi significativamente superior às combinações de modelos de referência que envolviam técnicas mais tradicionais como SVM e rede neural artificial (*Artificial Neural Network* - ANN) para grandes conjuntos de dados relacionados ao crédito.

Bequé e Lessmann [33], exploraram o potencial das máquinas de aprendizagem extrema (ELM), um tipo de rede neural artificial, para gerenciamento do risco de crédito ao consumidor. Os autores compararam empiricamente a ELM com técnicas de pontuação estabelecidas de acordo com três critérios de desempenho: facilidade de uso, consumo de recursos e precisão preditiva. Os resultados empíricos confirmaram as vantagens conceituais do ELM e indicam que eles são uma alternativa valiosa a outros métodos de modelagem de risco de crédito.

Wang, Chen [34] analisaram um conjunto de dados para avaliação de risco de crédito desbalanceado, no qual, compararam diferentes técnicas de correção: sobreamostragem sintética minonitária (SMOTE), subamostragem (*undersampling*), sobreamostragem (*oversampling*) e amostragem híbrida (*hybrid sampling*). Para construir classificadores de base mais adequados, funções de kernel múltiplas (Gaussiana, Polinomial e Sigmoide) são usadas respectivamente para melhorar o mapa auto-organizável fuzzy. Os resultados mostraram que, o novo modelo apresentado no artigo teve um desempenho melhor do que outros métodos em termos de avaliação do risco de crédito desbalanceado.

Sun e Lang [35] elaboraram um modelo de aprendizado de máquina baseado em árvore para avaliação de crédito em uma base de dados desbalanceada. Para tal, os autores utilizaram a técnica de SMOTE e no algoritmo de aprendizado de máquina com taxas de amostragem diferenciadas (DSR), que é denominado DTE-SBD (Conjunto de Árvores

de Decisão Baseado em SMOTE e Agregação e amostragem diferenciada) para corrigir o problema de desbalanceamento dos dados.

Durante o desenvolvimento do trabalho, identificou-se a necessidade de explorar técnicas adicionais que não haviam sido consideradas inicialmente e, por isso, não entraram na revisão bibliográfica da literatura. Devido à natureza dos dados, especialmente o desbalanceamento da variável resposta, tornou-se crucial adotar abordagens que levassem em consideração o custo dos erros nas previsões, penalizando de forma mais rigorosa os erros em classes de maior impacto. Com esse objetivo, foram incorporadas as técnicas de Naive Bayes com Custo e AdaBoost com Custo, que ajustam suas previsões de acordo com a gravidade dos erros, tornando os modelos mais adequados para lidar com o desbalanceamento e as consequências financeiras associadas [36][37].

Aprender a partir de dados desbalanceados apresenta desafios significativos para algoritmos de aprendizado de máquina, pois eles precisam lidar com a distribuição desigual de exemplos no conjunto de treinamento. Como os classificadores padrão tendem a ser inclinados para a classe majoritária, há uma necessidade de métodos específicos que possam superar essa dominância de uma única classe [38]. Naoki et al. [36], discorrem sobre um método para resolver problemas de aprendizagem sensível ao custo em múltiplas classes utilizando qualquer algoritmo de classificação binária. Destacam no entanto, que alterar o equilíbrio entre exemplos negativos e positivos no treinamento pode gerar pouco impacto nos classificadores gerados por métodos de aprendizado padrão, como os bayesianos e as árvores de decisão.

Qing-Yan et al. destacam que os algoritmos de potencialização (*boosting*) sensíveis ao custo têm se mostrado eficazes na resolução de problemas desafiadores relacionados ao desbalanceamento de classes. No entanto, a influência dos custos de classificação incorreta e do nível de desbalanceamento no desempenho dos algoritmos ainda não está completamente clara. No artigo desenvolvido, os autores realizam uma comparação empírica entre seis combinações de algoritmos que levam em consideração os custos, incluindo o AdaCost, que pertence à família do AdaBoost. Os resultados mostram que as combinações de algoritmos do AdaCost superam aquelas obtidas com a técnica de potencialização sensível ao custo (*Cost-Sensitive Boosting - CSB*) ao lidar com conjuntos de dados de diferentes níveis de desbalanceamento.

Capítulo 4

Fundamentação Teórica

Neste capítulo, será apresentada a fundamentação teórica dos conceitos mais relevantes para esta pesquisa. A escolha das referências foi baseada nos principais conceitos encontrados na revisão da literatura apresentada anteriormente, bem como em outras fontes relevantes para o estudo.

A Seção 4.1 aborda, de forma resumida, os conceitos de risco de crédito e ciclo de crédito. Em seguida, a Seção 4.2 descreve o conceito de dados abertos. A Seção 4.3 apresenta alguns modelos de classificação que serão utilizados neste trabalho. Por fim, a Seção 4.4 discute algumas técnicas para correção de desbalanceamento de dados.

4.1 Crédito

O Banco Central, em seu Artigo 21º da Resolução 4.577/2017 (BCB, 2017), define o risco de crédito como a possibilidade de ocorrência de perdas associadas a: I - o não cumprimento, pela contraparte, de suas obrigações nos termos pactuados; II - a desvalorização, redução de remunerações e ganhos esperados em instrumentos financeiros decorrentes da deterioração da qualidade creditícia da contraparte, do interveniente ou do instrumento mitigador; III - a reestruturação de instrumentos financeiros; ou IV - os custos de recuperação de exposições caracterizadas como ativos problemáticos [39].

A concessão de crédito é uma decisão tomada sob condições de incerteza. Seja em empréstimos, vendas a prazo ou prestação de serviços, independentemente de o crédito ser solicitado ou oferecido pelo credor, sempre existe a possibilidade de perda, o que é denominado risco de crédito [3].

Os modelos de risco de crédito podem ser utilizados em todo o ciclo de crédito, conforme resumido na Figura 4.1. Na concessão de crédito (*credit score)*, os modelos são usados para identificar a probabilidade de não pagamento e o valor a ser disponibilizado ao cliente. Na manutenção periódica (*behavior score)*, o uso dos modelos auxilia na definição do risco da operação, podendo resultar na extinção ou na expansão das linhas de crédito já fornecidas. Já nas etapas de cobrança e recuperação (*collection score)*, esses modelos são aplicados para definir estratégias de recebimento do crédito em atraso ou cessão do crédito em prejuízo [40].

Figura 4.1: Ciclo de Crédito [1].

Neste trabalho, o modelo desenvolvido será para a segunda etapa do ciclo de crédito: o escore de crédito, que é uma medida do risco de crédito. O modelo de escore de crédito é a denominação genérica dada no mercado às fórmulas de cálculo dos escores de crédito [3].

4.2 Dados abertos

O Portal Brasileiro de Dados Abertos foi desenvolvido pela sociedade e para a sociedade, sob a liderança do extinto Ministério do Planejamento. A política brasileira de dados abertos tem como objetivos fundamentais a promoção da transparência, o engajamento na participação social, o desenvolvimento de novos e melhores serviços governamentais, e o aumento da integridade pública. O fomento tecnológico por meio do uso de dados

abertos é o pilar principal para o desenvolvimento de governos mais acessíveis, eficazes e responsáveis [6].

Entre as informações disponíveis no portal estão os benefícios disponibilizados pelo Governo Federal aos cidadãos, incluindo o Auxílio Emergencial. O auxílio foi aprovado pelo Congresso Nacional em 2020, devido à pandemia do novo Coronavírus (Covid-19). O objetivo era garantir uma renda mínima aos brasileiros em situação mais vulnerável, uma vez que muitas atividades econômicas foram gravemente afetadas pela crise gerada pela pandemia [6].

4.3 Algoritmos de classificação

Nesta seção, serão apresentados alguns dos principais algoritmos de classificação utilizados na área de crédito e que serão modelados neste estudo.

4.3.1 Regressão Logística

A regressão logística é um modelo estatístico utilizado para prever a probabilidade de ocorrência de um evento binário, ou seja, um evento que pode assumir apenas dois valores possíveis, como sim/não ou verdadeiro/falso [2]. É uma técnica de aprendizado de máquina supervisionado e é comumente usada em problemas de classificação binária, como detecção de spam, análise de crédito, previsão de doenças, entre outros. Essa técnica é utilizada para estudar a relação entre uma variável categórica e um conjunto de variáveis explicativas [41].

Diferentemente da regressão linear, que é usada para prever uma variável contínua, a regressão logística produz uma saída binária, ou seja, uma probabilidade entre 0 e 1 de que um evento ocorra [1]. Essa probabilidade é então mapeada para uma classe binária, geralmente usando um limiar de 0,5. Se a probabilidade calculada for maior que o limiar, a classe prevista é "1"; caso contrário, é "0".

A regressão logística utiliza a função sigmóide para modelar a relação entre as variáveis independentes e a variável dependente. Essa função possui um formato em S, conforme mostra a Figura 4.2, que transforma a soma ponderada das variáveis independentes em uma probabilidade entre 0 e 1. A equação da função logística é dada por [2][1]:

$$\text{Função logística:} \quad p = \frac{1}{(1 + \exp^{-a})} \tag{4.1}$$

Onde:

- p é a probabilidade condicional da variável dependente assumir o valor 1 (o evento binário ocorrer).
- a é a soma ponderada das variáveis independentes, dada pela fórmula:

$$\text{Soma ponderada:} \quad b_0 + b_1x_1 + b_2x_2 + ... + b_kx_k, \tag{4.2}$$

onde $b_0, b_1, b_2, ..., b_k$ são os coeficientes do modelo e $x_1, x_2, ..., x_k$ são as variáveis independentes.

Por sua vez, a função *logit* é a inversa da função logística e é usada para modelar a relação entre as variáveis independentes e a *log-odds* da variável dependente assumir o valor 1. A *log-odds* é a transformação logarítmica da razão entre a probabilidade da variável dependente assumir o valor 1 e a probabilidade de assumir o valor 0. A equação da função logit é dada por:

$$\text{Função logit:} \quad \ln\left(\frac{p}{1-p}\right) = \ln(p) - \ln(1-p) = a \tag{4.3}$$

Onde:

- ln é o logaritmo natural.
- p é a probabilidade condicional da variável dependente assumir o valor 1.
- $1 - p$ é a probabilidade condicional da variável dependente assumir o valor 0.

A partir da equação 4.3, verifica-se que quando $p = 0.5$, a função será igual a 0. Substituindo o valor de p na função 4.4:

$$\text{Exemplo de função Logit:} \quad \ln\left(\frac{0.5}{0.5}\right) = \ln(1) = 0 \tag{4.4}$$

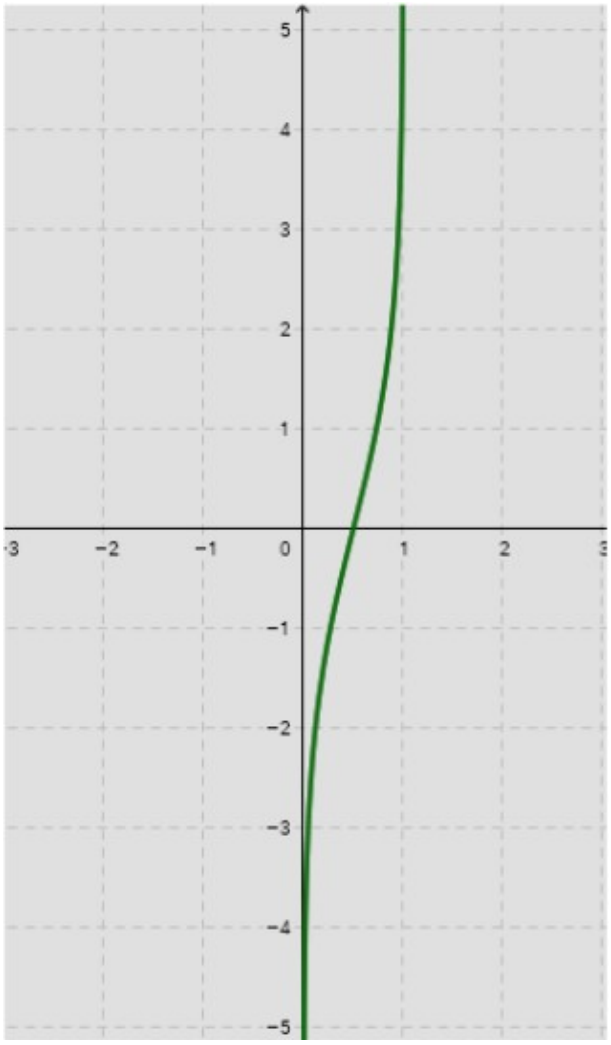


Figura 4.2: Função Logit [2].

Aplicando o inverso da função logit 4.4 temos [2]:

$$\text{Inversa da função logit:} \quad \text{logit}^{-1}(\alpha) = \frac{1}{1+e^{-\alpha}} = \frac{e^{\alpha}}{1+e^{\alpha}} \tag{4.5}$$

Onde α é a combinação linear das variáveis e seus coeficientes. A inversa da função *logit* retorna a probabilidade da variável dependente y ser igual a 1 (caso de sucesso). Na Figura 4.3, observa-se que o gráfico da inversa do logit é o mesmo do logit, apenas rotacionado 90 graus. Basicamente, houve uma troca das coordenadas x e y, agora com o domínio da função de 0 a 1 no eixo y, resultando no formato de sigmóide mencionado anteriormente.

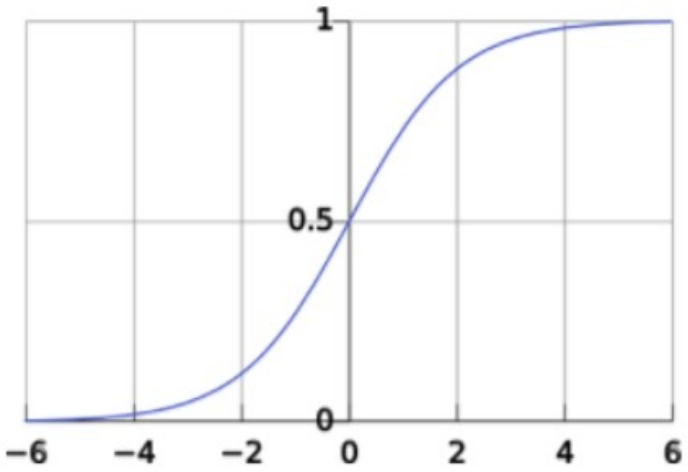


Figura 4.3: Curva Logística [1].

Na regressão logística, a estimação dos parâmetros não pode ser feita por mínimos quadrados, pois a variável dependente é binária. Assim, a estimação é realizada por máxima verossimilhança. A probabilidade para quando $y_i = 1$ é $P_r(y_i = 1|x_i) = \pi(x_i)$, e para $y_i = 0$ é $P_r(y_i = 0|x_i) = 1 - \pi(x_i)$. Logo, para cada indivíduo, a função de verossimilhança é dada por [1]:

$$\text{Verossimilhança:} \quad L_i = \pi(x_i)^{y_i}[1 - \pi(x_i)]^{(1-y_i)} \tag{4.6}$$

Para os valores de $y_i = 1$ ou 0, para todo i variando de 1 a n.

As observações são independentes, então a função de verossimilhança total pode ser escrita como:

$$\text{Verossimilhança Total:} \quad L = \prod_{i=1}^{n} \pi\left(x_i\right)^{y_i}\left[1 - \pi(x_i)\right]^{(1-y_i)} \tag{4.7}$$

Para encontrar o ponto máximo da função de verossimilhança, é necessário o uso de métodos iterativos, como o método de Newton-Raphson, que inicia com valores arbitrários e, avaliando os erros de previsão, gera uma sequência de soluções até convergir para a solução que maximiza a função [1].

4.3.2 Florestas Aleatórias

Modelos baseados em árvores, introduzidos por Leo Breiman na década de 1980, são amplamente aplicados tanto em tarefas de classificação quanto em regressão. Esses modelos são populares por sua simplicidade conceitual e computacional, além de sua inter-

pretabilidade. No entanto, modelos simples baseados em árvores geralmente apresentam uma precisão inferior em comparação com métodos de aprendizado mais complexos. Para superar essa limitação, métodos generalizados, como as Florestas Aleatórias *(Random Forests)*, foram desenvolvidos, combinando a simplicidade das árvores de decisão com a robustez obtida a partir da combinação de múltiplas árvores, gerando previsões mais precisas. Embora eficazes, uma das desvantagens dessa técnica é a dificuldade de interpretação dos resultados [42].

O princípio central da técnica de Florestas Aleatórias é a construção de múltiplas árvores de decisão independentes, combinadas para gerar previsões mais robustas e precisas. Essas árvores são criadas a partir de subconjuntos aleatórios dos dados de treinamento e subconjuntos aleatórios das variáveis preditoras. Essa aleatoriedade na construção de cada árvore reduz o risco de sobreajuste *(overfitting)* e garante que cada árvore contribua com uma perspectiva única para a previsão final. No contexto de classificação, a floresta decide pela classe mais votada, enquanto em regressões a previsão final é dada pela média das saídas das árvores [43]. Florestas Aleatórias têm sido amplamente utilizadas em domínios como finanças, marketing, saúde e ciência de dados, devido à sua robustez e capacidade de generalização [44].

Entre as principais vantagens dessa técnica está sua habilidade de lidar com dados ausentes, além de ser adaptável tanto para variáveis categóricas quanto numéricas, e operar bem em conjuntos de dados desbalanceados. Outra vantagem é a ausência de necessidade de pré-processamento extensivo, o que facilita sua implementação prática. No entanto, apesar de sua flexibilidade, Florestas Aleatórias podem ser computacionalmente custosas, tanto no treinamento quanto na previsão. Além disso, sua interpretabilidade é um desafio, pois a aleatoriedade na construção das árvores torna mais complexa a identificação de quais variáveis são decisivas para o modelo. A Figura 4.4 a seguir mostra como as florestas aleatórias chegam à sua previsão final combinando os resultados das árvores de decisão individuais.

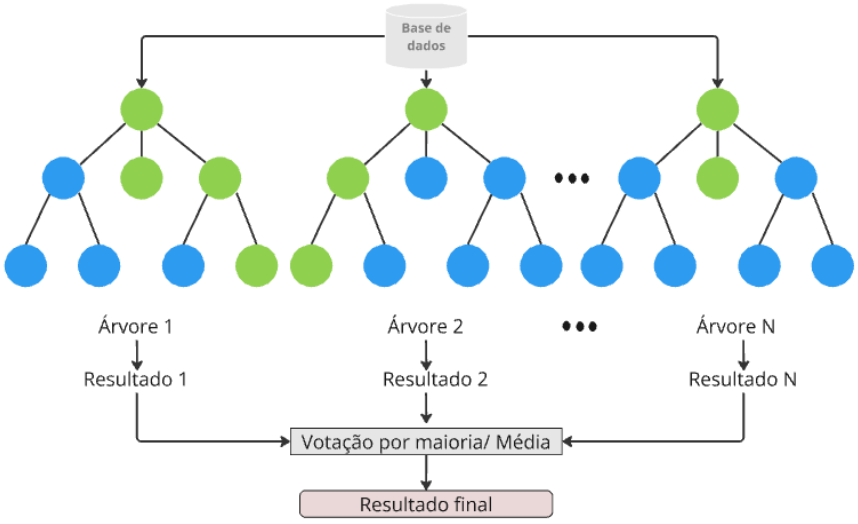


Figura 4.4: Floresta Aleatória - elaborado pelo autor.

No caso de um problema de classificação, cada árvore realiza uma votação para a classe que prevê ser correta, e a classe que receber mais votos é a classe final predita pelo modelo. Já para regressão, a previsão final é a média das saídas de todas as árvores. A vantagem desse processo é a redução de viés e de variabilidade em relação ao uso de uma única árvore, o que aumenta a robustez do modelo. Esta combinação de múltiplas árvores permite que o modelo seja mais resistente ao sobreajuste e tenha um desempenho superior em relação a uma única árvore de decisão.

4.3.3 LightGBM

A técnica Máquina de Aprendizado de Gradiente Acelerado por Luz *(Light Gradient Boosting Machine, LightGBM)* foi proposta por Guolin Ke em 2017, e é um algoritmo baseado em árvores amplamente utilizado para resolver problemas de classificação e regressão. Uma das suas principais características é a eficiência com grandes conjuntos de dados, devido ao seu algoritmo que utiliza impulsionamento de gradiente *(Gradient Boosting)*. O LightGBM diferencia-se ao empregar o crescimento baseado em folhas *(Leaf-wise Growth)*, que prioriza a divisão das folhas que mais reduzem o erro, em vez de crescer a árvore de maneira simétrica, como fazem os métodos tradicionais [45].

Estudos mostram que o LightGBM frequentemente supera o impulsionamento de gradiente *(Gradient Boosting)* tradicional, ao mesmo tempo que consome menos recursos computacionais [32]. Outra vantagem é sua capacidade de lidar com desbalanceamento de classe em problemas de classificação. Ele atribui pesos diferentes às amostras, o que

melhora a precisão em classes minoritárias. Além disso, o LightGBM oferece diversos métodos de ajuste de hiperparâmetros, como busca aleatória *(random search)* e otimização Bayesiana, para otimizar o desempenho do modelo [45][46].

Embora a ideia central do impulsionamento de gradiente com árvores de decisão *(Gradient Boosting Decision Tree)* e do LightGBM seja similar, o LightGBM tem um desempenho superior devido a duas inovações-chave: o crescimento foliar *(Leaf-wise Growth)* e o uso de histogramas para busca de divisões. A abordagem foliar permite que o LightGBM cresça as árvores dividindo as folhas que mais reduzem o erro, o que pode acelerar a redução da perda, mas aumenta o risco de sobreajuste se não for controlado adequadamente [32].

4.3.4 Modelos com custo

As técnicas de Naive Bayes com Custo e AdaBoost com Custo se destacam por sua capacidade de incorporar uma matriz de custo, permitindo que o modelo não apenas maximize a acurácia geral, mas também ajuste suas previsões de acordo com os impactos específicos dos erros em diferentes classes[38][19]. Ao contrário de métodos tradicionais que tratam todos os erros de forma igual, essas técnicas penalizam mais fortemente erros com consequências maiores, como os falsos negativos no contexto de inadimplência, por exemplo [47].

Essa abordagem torna os modelos mais sensíveis à detecção de clientes que podem ficar inadimplentes, priorizando a minimização de perdas financeiras em detrimento de acertos em classes de menor relevância. Diferentemente de outras técnicas de classificação testadas neste trabalho, como árvores de decisão ou regressão logística, que visam otimizar métricas globais como acurácia ou precisão, essas abordagens baseadas em custo moldam suas previsões para otimizar o impacto financeiro ou operacional, oferecendo uma solução mais personalizada para problemas com riscos desiguais entre as classes.

Naive Bayes

A teoria de decisão bayesiana é uma abordagem estatística fundamental para o problema de classificação de padrões. O objetivo é quantificar a relação custo-benefício entre várias decisões de classificação utilizando probabilidades e os custos que acompanham essas decisões. Nesse contexto sensível a custos, pode-se escolher um algoritmo de aprendizado e otimizar seu desempenho ajustando não apenas seus parâmetros, mas também os custos de classificação incorreta [48].

O Naive Bayes Sensível ao Custo (CSNB) é uma adaptação do algoritmo Naive Bayes para lidar com problemas onde diferentes erros de classificação têm custos distintos. Enquanto o Naive Bayes convencional visa maximizar a probabilidade de acerto da classificação sem considerar os custos associados, o CSNB incorpora os custos de erros para tomar decisões que minimizam o custo esperado, em vez de apenas maximizar a precisão [47].

O CSNB ajusta o processo de decisão ao introduzir uma função de custo que penaliza mais fortemente certos tipos de erros de classificação. Isso é feito através de uma matriz de custo [47] $L(C_k, C_j)$, onde:

- $L(C_k, C_j)$ é o custo de classificar uma instância C_k quando a classe verdadeira é C_j.

Assim, ao invés de escolher a classe C_k que maximiza a probabilidade a posteriori, o CSNB seleciona a classe que minimiza o custo esperado da classificação. Logo, a função de decisão será dada por:

$$\hat{C} = arg\frac{min}{C_k} \sum_{C_j} P\left(C_j \middle| X\right).L(C_k, C_j) \tag{4.8}$$

onde:

- $P\left(C_j \middle| X\right)$ é a probabilidade posterior da classe C_j (calculada de acordo com o Teorema de Bayes).
- $L(C_k, C_j)$ é o custo de classificar como C_k quando a classe verdadeira é C_j.

AdaBoost

A técnica de Reforço Adaptativo (Adaptive Boosting - AdaBoost) é um método de aprendizagem supervisionada que visa melhorar o desempenho de classificadores fracos, ou seja, classificadores com desempenho ligeiramente melhor que o acaso. Ela faz isso combinando diversos classificadores fracos em um modelo forte, com ênfase em corrigir os erros cometidos nas iterações anteriores [49].

Na técnica, inicialmente, cada amostra no conjunto de treinamento recebe um peso igual. Para N amostras, o peso inicial de cada amostra é $1/N$. No momento do treino do classificador tido como fraco, o treino ocorre considerando o conjunto de treinamento ponderado. O erro do classificador fraco é calculado como a soma dos pesos das instâncias mal classificadas:

$$\epsilon = \sum_{i=1}^{N} .I(y_i \neq h(x_i)) \tag{4.9}$$

onde w_i é o peso da amostra i, y_i é o rótulo verdadeiro da amostra i, $h(x_i)$ é a predição do classificador, e I(.) é a função indicadora, que é 1 quando a predição está errada e 0 quando está correta. Em seguida, é necessário realizar o ajuste dos pesos do classificador. Um peso α_t é atribuído ao classificador fraco baseado em seu desempenho:

$$\alpha = \frac{1}{2} ln(\frac{1 - \epsilon_t}{\epsilon_t}) \tag{4.10}$$

Quanto menor o erro ϵ_t, maior será o peso α_t, indicando maior confiança no classificador fraco. Os pesos das amostras são atualizados para dar mais ênfase às instâncias mal classificadas:

$$w_i^{(t+1)} = w_i t.e^{\alpha_t . I(y_i \neq h_t(x_i))} \tag{4.11}$$

Isso significa que instâncias mal classificadas têm seus pesos aumentados, enquanto as bem classificadas têm seus pesos diminuídos. O processo de treinamento do próximo classificador fraco é repetido, levando em consideração os novos pesos das amostras. O processo continua por várias iterações até que o número máximo de classificadores fracos seja alcançado ou até que o erro global seja suficientemente baixo. A predição final do AdaBoost é feita pela combinação ponderada dos classificadores fracos:

$$H(x) = sign(\sum_{t=1}^{T} \alpha_t . h_t(x)) \tag{4.12}$$

onde $h_t(x)$ é a predição do classificador fraco t e α_t é o seu peso.

Ao considerar o custo, a ideia principal é modificar o algoritmo original para considerar os custos de diferentes tipos de erros de classificação. Isso significa que ao invés de tratar todos os erros de classificação da mesma forma, o AdaBoost com Custo usa uma matriz de custos $L(C_i, C_j)$, onde se tem o custo de classificar uma instância da classe verdadeira C_i como C_j.

Isso significa que, ao invés de aumentar os pesos das instâncias mal classificadas de maneira uniforme, o algoritmo ajusta os pesos levando em consideração o custo do erro. Por exemplo, um erro que tenha um custo maior receberá um aumento maior no peso da instância correspondente. A fórmula de atualização de pesos pode ser ajustada como:

$$w_i^{(t+1)} = w_i t.e^{\alpha_t . L(y_i h_t(x_i))} \tag{4.13}$$

onde, $L(y_i h_t(x_i))$ é o custo de classificar erroneamente a instância i. A combinação final dos classificadores fracos leva em consideração tanto os pesos dos classificadores quanto os custos associados aos erros de classificação será dado por:

$$H(x) = sign(\sum_{t=1}^{T} \alpha_t.h_t(x).L(y_i, h_t(x))) \tag{4.14}$$

Isso ajusta a decisão final de forma a minimizar os custos totais de classificação errada [49].

4.3.5 Técnicas de correção de desbalanceamento

O desbalanceamento de classes é um problema comum em modelos de crédito, pois, em geral, a maioria das solicitações é aprovada, enquanto apenas uma pequena fração é negada. Isso pode introduzir um viés no modelo, fazendo com que ele favoreça as classes majoritárias em detrimento das minoritárias. Embora muitos modelos estatísticos e de inteligência artificial tenham sido amplamente estudados para a avaliação de crédito, a maioria se baseia em conjuntos de dados balanceados, e a modelagem de crédito em cenários desbalanceados ainda carece de estudos mais aprofundados. No mundo econômico real, empresas de alto risco são menos numerosas do que as empresas comuns, o que torna a avaliação de crédito um problema típico de classificação desbalanceada [35].

A avaliação do risco de crédito é um problema multidimensional e que lida com dados desbalanceados, baseado principalmente em um grande volume de dados históricos, como: status de emprego, histórico de crédito, status da conta pessoal, entre outros. O uso de todos esses recursos pode, por um lado, aumentar a cobertura, mas, por outro, diminuir a precisão do modelo, exigindo, assim, uma abordagem de seleção de variáveis para lidar adequadamente com dados desbalanceados [9].

Existem algumas abordagens que buscam corrigir esse problema, não apenas em modelos de crédito, mas também em algoritmos de aprendizado de máquina. De maneira geral, a ideia é modificar a distribuição dos dados por meio da reamostragem, que envolve aumentar a quantidade de dados da classe minoritária ou reduzir a quantidade da classe majoritária. A reamostragem pode ser dividida em duas categorias: sobreamostragem (*oversampling*), que consiste em aumentar a classe minoritária, e subamostragem (*undersampling*), que reduz a classe majoritária.

A técnica de sobreamostragem consiste em aumentar o número de amostras da classe minoritária. Isso pode ser feito através da replicação de amostras existentes ou da geração de novas amostras sintéticas usando técnicas de amostragem aleatória, como a *SMOTE (Synthetic Minority Over-sampling Technique)* que usa k-vizinhos mais próximos para produzir novas instâncias com base na distância entre os dados da minoria e alguns vizinhos

mais próximos selecionados aleatoriamente. Por outro lado, a técnica de *undersampling* consiste em reduzir o número de amostras da classe majoritária [9]. A abordagem de *undersampling* tem a desvantagem de que é fácil perder as informações importantes ao descartar amostras aleatoriamente. Contudo, essa abordagem é mais adequada para lidar com o conjunto de dados com grande tamanho de amostra devido à economia de tempo de computação [34].

É possível ainda combinar as técnicas de sobreamostragem e subamostragem. Essa abordagem aumenta o número de classes minoritárias por meio da sobreamostragem, diminui o número de classes majoritárias por meio da subamostragem e, por fim, obtém o conjunto de dados balanceado. Devido à combinação, a abordagem de amostragem híbrida pode reduzir o grau de sobreajuste (*overfitting*). Além disso, ela tem uma estrutura simples e tempo de computação rápido [34].

Para auxiliar na compreensão do desbalanceamento dos dados é muito comum a utilização de métricas de avaliação. Ao lidar com classes desbalanceadas, é importante usar métricas como precisão, sensibilidade, F1-score e área sob a curva ROC (AUC-ROC) que são mais informativas do que a acurácia, pois levam em consideração o desequilíbrio dessas classes.

4.4 Métricas de avaliação do modelo

A validação final dos modelos de crédito consiste em duas atividades básicas: avaliação da fórmula por analistas de crédito e a avaliação baseada em critérios estatísticos [3]. Neste trabalho, serão apresentados alguns critérios estatísticos para avaliação de modelos.

Na área de crédito, existem diversas métricas que podem ser utilizadas para avaliar a performance de modelos de crédito. Parte dessas métricas são oriundas da matriz de confusão. A matriz de confusão Figura 4.5 é uma tabela que organiza as previsões do modelo em relação aos valores verdadeiros em quatro categorias diferentes: verdadeiro positivo (TP), falso positivo (FP), verdadeiro negativo (TN) e falso negativo (FN) [9][34][35]. Essas categorias são derivadas da comparação entre a previsão do modelo e o valor real pagamento do crédito concedido.

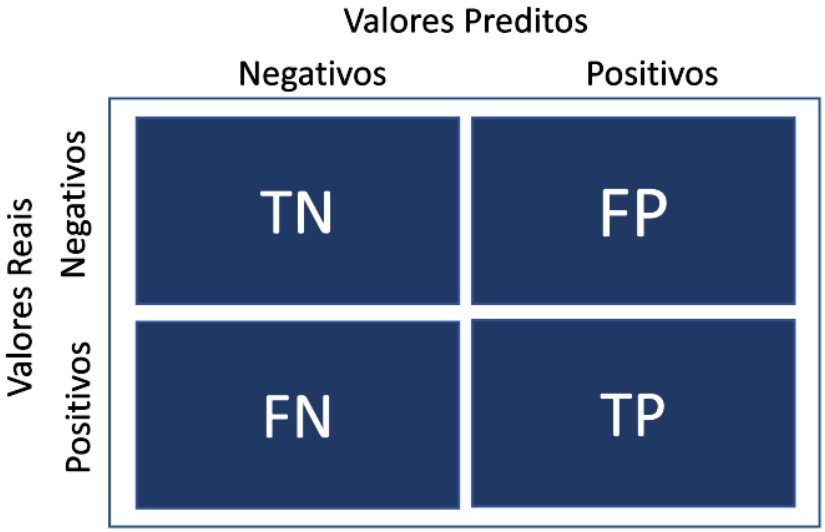


Figura 4.5: Matriz de confusão - elaborado pelo autor.

A partir da combinação das categorias acima mencionadas, algumas métricas podem ser calculadas:

- **Precisão**: Mede a proporção de verdadeiros positivos em relação a todas as instâncias classificadas como positivas pelo modelo:

$$\frac{TP}{TP + FP} \tag{4.15}$$

Uma alta precisão indica que há uma baixa taxa de falsos positivos, o que é desejável para evitar concessões de crédito inadequadas [9][34][35].

- **Revocação ou sensibilidade**: Mede a proporção de verdadeiros positivos em relação a todas as instâncias verdadeiramente positivas:

$$\frac{TP}{TP + FN} \tag{4.16}$$

Uma alta revocação indica que o modelo é capaz de identificar a maioria dos bons pagadores [9][34][35].

- **Especificidade**: Mede a proporção de verdadeiros negativos em relação a todas as instâncias verdadeiramente negativas:

$$\frac{TN}{TN + FP} \tag{4.17}$$

Uma alta especificidade indica que o modelo é capaz de identificar a maioria dos maus pagadores [9][34][35].

- **Acurácia**: É a proporção de previsões corretas em relação ao total de previsões:

$$\frac{TP + TN}{TP + FP + TN + FN} \tag{4.18}$$

Em geral, uma alta acurácia indica um bom desempenho do modelo. No entanto, em se tratando de dados desbalanceados a acurácia não pode ser a única métrica de avaliação porque não considera que os falsos positivos são mais importantes do que os falsos negativos [9].

Outras métricas comumente utilizadas na área de crédito são:

- **G-mean**: É calculada pela sensibilidade e especificidade ao mesmo tempo e pode ser usada para avaliar o equilíbrio entre o desempenho da classificação entre as classes minoritárias e majoritárias [9]:

$$\sqrt{Sensibilidade * Especificidade} \tag{4.19}$$

- **KS**: O índice de Kolmogorov-Smirnov (KS) é usado como uma métrica para avaliar a qualidade e o poder preditivo dos modelos de *credit scoring*. O KS mede a capacidade do modelo de distinguir entre os bons e os maus pagadores com base nas características dos solicitantes de crédito.

  Para calcular o KS em crédito, os solicitantes são ordenados de forma decrescente de acordo com o escore atribuído pelo modelo. Em seguida, a taxa acumulada de bons pagadores e a taxa acumulada de maus pagadores são calculadas. O KS é a diferença máxima entre essas duas taxas acumuladas.

  Um alto valor de KS indica que o modelo tem uma boa capacidade de discriminação entre bons e maus pagadores. Por outro lado, um valor baixo de KS indica que o modelo não é capaz de distinguir efetivamente entre esses grupos. Adicionalmente, esse medida é frequentemente utilizada para avaliar e comparar diferentes modelos de avaliação de crédito. Os modelos com KS mais altos são considerados mais eficazes na identificação de clientes com maior probabilidade de inadimplência [3].

- **AUROC - Area-Under Receiver Operating Characteristic**: A curva ROC *(Receiver Operating Characteristic)* baseia-se em duas definições: especificidade e sensibilidade. A sensibilidade/sensitividade pode ser entendida como a capacidade de identificar os bons clientes; a especificidade pode ser entendida como a capacidade

de identificar os maus clientes [9][3]. A área sob a curva ROC é denotada por AUROC *(Área Under ROC)*. Para obter a curva, conforme mostra a Figura 4.6, calcula-se a sensitividade e a especificidade para cada valor, variando do menor para maior *score*.

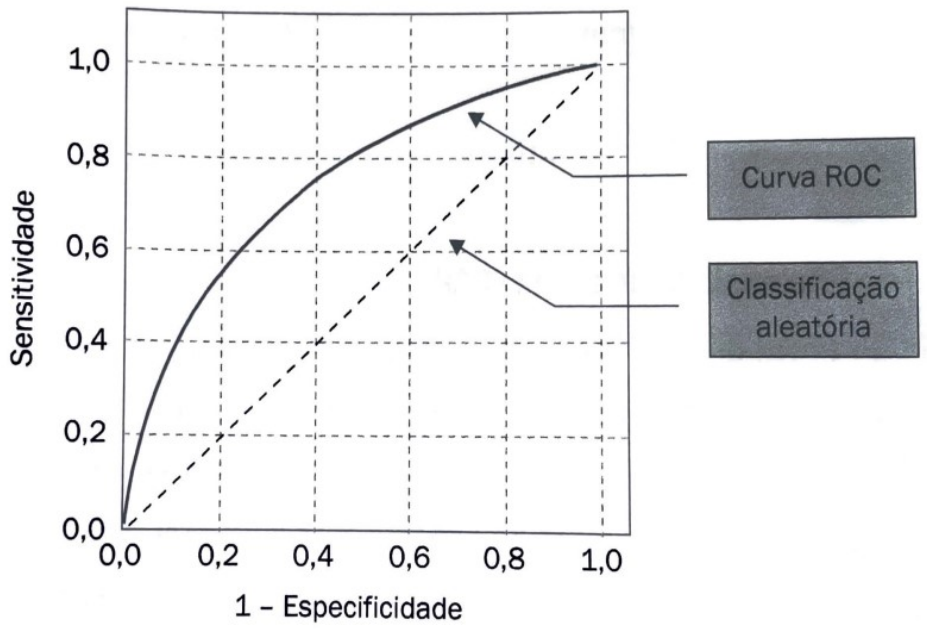


Figura 4.6: Curva ROC [3].

- **Kappa**: O Coeficiente Kappa (ou Kappa de Cohen) é uma métrica estatística utilizada para medir o grau de concordância entre dois avaliadores ou classificadores sobre um conjunto de itens, levando em consideração o acaso. Diferente de outras métricas de acurácia, o Kappa corrige a probabilidade de concordância que ocorreria ao acaso, tornando-o uma medida mais robusta para avaliar a precisão de classificações categóricas. [3]. Os valores de Kappa (k) podem variar de -1 a 1, onde k=1 significa uma concordância perfeita, k=0 significa uma concordância equivalente ao acaso e k<0 seria uma concordância pior do que o acaso, indicando uma possível discordância sistemática [50].

- **K-fold**: O K-Fold Validação Cruzada é uma técnica de validação utilizada para avaliar o desempenho de modelos de aprendizado de máquina. Ele funciona dividindo o conjunto de dados em K partes (ou folds) de tamanho aproximadamente igual. Em cada iteração, um dos *folds* é reservado para teste (conjunto de validação), enquanto os outros K-1 *folds* são usados para treinar o modelo. O processo se repete K vezes, de modo que, ao final, cada *fold* tenha sido usado exatamente uma vez como conjunto de teste. A Figura 4.7, resume este processo. Por fim, a média

dos resultados obtidos em cada iteração é utilizada para avaliar a performance final do modelo. Como o modelo é avaliado em várias iterações com diferentes porções de dados, o K-Fold ajuda a reduzir o risco de sobreajuste ao testar o modelo em diferentes amostras dos dados.[3][42].

Figura 4.7: K-fold - elaborado pelo autor.

- **Otimização Baseada em Tarefas**: Optuna é uma técnica eficiente de otimização de hiperparâmetros projetada para automatizar o processo de ajuste de modelos de aprendizado de máquinas. Seu funcionamento é baseado em buscas *bayesianas* e oferece uma interface simples, mas poderosa, para encontrar a combinação ideal de parâmetros. Além disso, o Optuna oferece um recurso chamado pode (*pruning*), que interrompe execuções de experimentos pouco promissores, economizando tempo e recursos computacionais. Sua flexibilidade e fácil integração com diferentes técnicas de aprendizado de máquinas, o tornam uma ferramenta poderosa para ajuste de hiperparâmetros, especialmente em cenários com múltiplas técnicas de balanceamento de dados e diferentes métricas de avaliação [51].

- **CatBoost**: O *CatBoost* é um algoritmo de aprimoramento por gradiente que constrói árvores de decisão em uma sequência, onde cada nova árvore corrige os erros das árvores anteriores. Ele utiliza uma abordagem que minimiza a função de perda, otimizando a predição com base nas previsões anteriores. Essa técnica utiliza técnicas de codificação para lidar com variáveis categóricas. O algoritmo transforma essas variáveis em números, preservando a ordem e a relação entre os valores sem a necessidade de extensivas manipulações de dados [42].

Capítulo 5

Estudo de Caso

O estudo de caso deste trabalho tem como objetivo testar as técnicas de aprendizado de máquina mencionadas no Capítulo 4, além de incorporar os dados do auxílio emergencial em um modelo de concessão de crédito, a fim de verificar o impacto dessas informações na capacidade preditiva do modelo. Para isso, será utilizado o método CRISP-DM, que é composto por seis etapas, descritas a seguir:

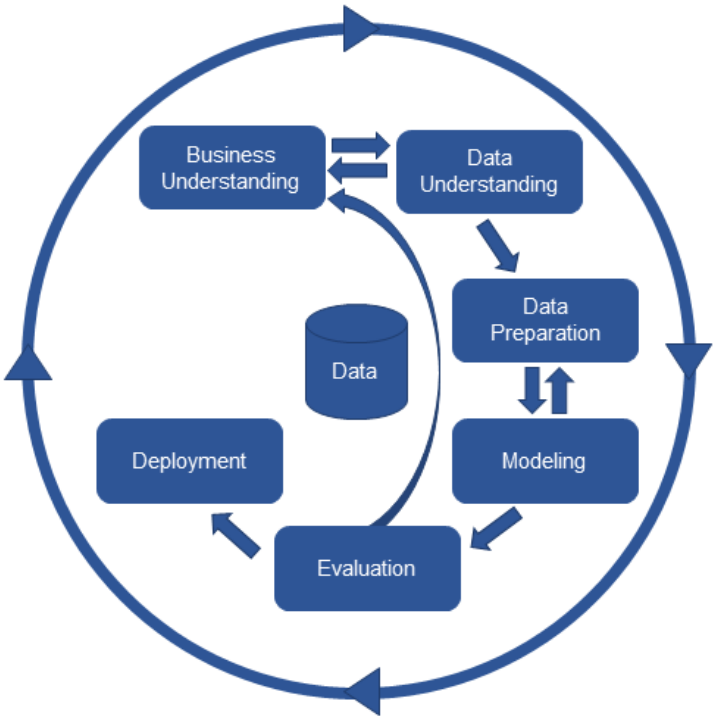


Figura 5.1: CRISP-DM [4].

O Cross Industry Standard Process for Data Mining (CRISP-DM) é um modelo de processo amplamente utilizado tanto no meio acadêmico quanto no mercado para mine-

ração de dados. O CRISP-DM é considerado um padrão completo, que visa facilitar a compreensão e a revisão sistemática das etapas envolvidas na solução de um problema, como proposto neste estudo [4].

1. **Entendimento do negócio** (*Business Understanding)* - entender do negócio e dos seus respectivos objetivos, definindo o problema e um plano preliminar para alcançar os resultados;
2. **Entendimento dos dados** (*Data Understanding)* - consiste na etapa de coletar os dados e compreendê-los, identificar problemas, buscar informações ou subconjuntos relevantes para revelar informações ocultas;
3. **Preparação dos dados** (*Data Preparation)* - consiste nas atividades para construir o conjunto de dados final para análise;
4. **Modelagem** (*Modeling)* - momento em que as técnicas são selecionadas e aplicadas, e seus parâmetros são calibrados;
5. **Avaliação** (*Evaluation)* - o modelo obtido é avaliado e revisado minuciosamente para que atenda aos objetivos;
6. **Implantação** (*Deployment)* - isto normalmente não é o fim do projeto. Mesmo que o objetivo do modelo seja aumentar o conhecimento dos dados, o conhecimento adquirido precisará ser organizado e apresentado de forma útil ao usuário/cliente.

Ressalta-se que, neste trabalho, iremos avançar até a fase de avaliação do modelo.

5.1 Entendimento do Negócio

A instituição financeira em questão tem como propósito oferecer serviços financeiros acessíveis e inclusivos a pessoas de baixa renda ou sem acesso aos serviços bancários tradicionais. Sua missão é promover a inclusão financeira e melhorar a qualidade de vida dessas pessoas por meio de soluções digitais e financeiras inovadoras.

A empresa oferece uma variedade de produtos e serviços financeiros, como contas digitais, cartões pré-pagos e empréstimos, sendo este último o foco deste estudo, além de outros serviços bancários básicos. Através de parcerias estratégicas com varejistas e outras empresas, a instituição busca facilitar o acesso de populações vulneráveis a serviços financeiros, permitindo que essas pessoas realizem transações, recebam pagamentos, economizem e planejem suas finanças de maneira mais eficiente. O produto de empréstimo pessoal foi lançado em junho de 2021, com o objetivo de proporcionar uma experiência

financeira inclusiva, transparente e conveniente, adaptada às necessidades e realidades dessas populações.

5.2 Entendimento dos Dados

A base de dados utilizada foi gerada a partir da junção de diferentes tabelas armazenadas no diretório da instituição, e todo o processamento, bem como o modelo desenvolvido, foi feito no ambiente *Databricks* [52]. A base de dados final conta com um total de 634.510 clientes distribuídos nos seguintes meses, conforme mostra 5.1:

Tabela 5.1: Distribuição de clientes por safras de desenvolvimento no modelo

Safra	**Quantidade**
202109	204.979
202110	211.237
202111	218.294
Total	634.510

Destaca-se que as safras acima citadas foram construídas com uma janela de 3 meses das informações. A Tabela 5.1 mostra a janela dos dados que foram considerados em cada uma das safras que serão utilizadas no desenvolvimento do modelo.

Explorando um pouco os dados, pode-se ter uma noção do perfil dos clientes que compõe a base de dados. Nota-se, a partir da Figura 5.2, que os clientes tem, em sua maioria, entre 40 e 60 anos de idade.

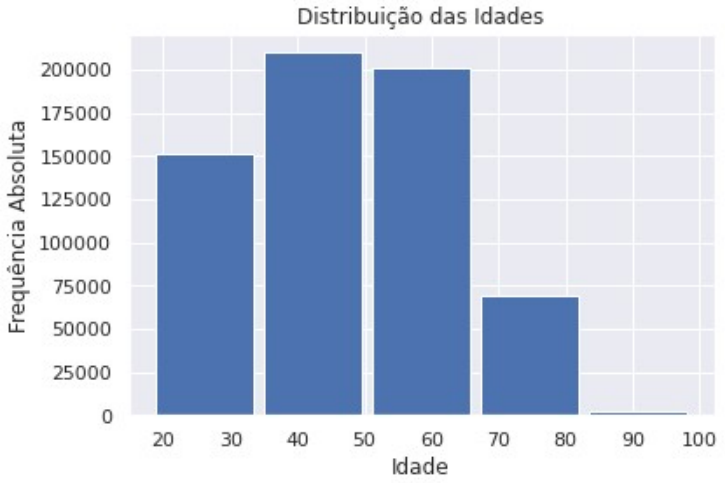


Figura 5.2: Distribuição das Idades.

Outra informação explorada diz respeito ao tempo de relacionamento com a instituição, onde a maioria dos clientes possui até 15 meses de conta, como pode ser observado na Figura 5.2.

Figura 5.3: Tempo de conta em meses.

É interessante verificar também a distribuição dos clientes por região declarada no momento do cadastro (Figura 5.4). Grande parte está concentrada na região sudeste do país.

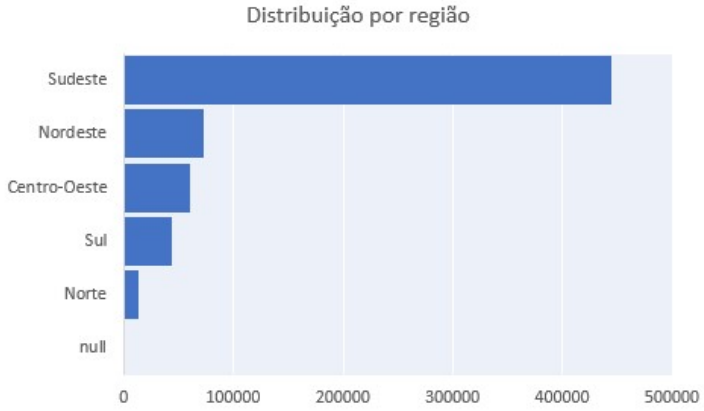


Figura 5.4: Distribuição dos clientes por região.

Essa análise inicial é importante para compreender se a carteira de clientes está aderente a proposta de democratizar o crédito.

5.3 Preparação dos dados

Para criação das possíveis variáveis explicativas foram consideradas informações cadastrais dos clientes, respeitando a Lei Geral de Proteção de Dados (LGPD), assim como, informações sobre transações, saldos, compras, pagamentos e interesse de contratação do Empréstimo Pessoal (EP).

Em acordo com a instituição financeira, as variáveis utilizadas neste trabalho não serão apresentadas com 100% de detalhamento. Será possível fornecer uma descrição geral das variáveis que compõem o modelo, destacando suas categorias e funções principais. No entanto, alguns detalhes específicos foram restringidos para resguardar a segurança da informação e evitar qualquer exposição que possa comprometer os dados sensíveis da instituição ou dos clientes envolvidos.

Todas as informações foram processadas, como dito anteriormente, no *Databricks*, sendo que, a maioria das variáveis foram criadas em *Structured Query Language* (SQL) por se entender que eram combinações relativamente simples. Destaca-se contudo que, outras variáveis e todo o desenvolvimento do modelo foi realizado utilizando a linguagem *python*. Inicialmente foram geradas 67 variáveis, listadas na Tabela 5.2 a seguir:

Tabela 5.2: Descrição das variáveis do modelo

Descrição	**Tipo do Dado**
Mês e ano das informações	Numérico
Variável resposta	Dicotômica
Cliente inativo com 30 e 90 dias	Dicotômica
Cliente inativo com 90 dias	Dicotômica
Cliente inativo com 30 dias	Dicotômica
Idade dos clientes	Numérico
Domínio do e-mail cadastrado	Categórica nominal
No e-mail cadastrado possui o primeiro nome	Dicotômica
Possui número na composição do e-mail cadastrado	Dicotômica
Percentual de número em relação a todos os caracteres	Numérico
Tempo de relacionamento do cliente	Numérico
Cliente ativo nos últimos 30 dias	Numérico
Cliente ativo nos últimos 60 dias	Numérico
Cliente ativo nos últimos 90 dias	Numérico
Unidade federativa do cliente a partir do CPF	Categórica nominal
Continua na próxima página	

Tabela 5.2 – continuação da página anterior

Descrição	Tipo do Dado
Total de pagamentos cdc realizados pelo cliente	Numérico
Total de pagamentos de boletos	Numérico
Total de pix que entrou na conta do cliente	Numérico
Total de cartão físico	Numérico
Total de cartão virtual	Numérico
Total de contas que foram pagas	Numérico
Valor total dos boletos	Numérico
Total de entradas	Numérico
Total de saídas	Numérico
Saldo	Numérico
Valor total em compras do cliente	Numérico
Quantidade total de compras	Numérico
Saldo Inicial	Numérico
Máximo de parcelas realizadas	Numérico
Quantidade total de contratos	Numérico
Setor predominante	Categórica nominal
Diretoria realizada	Categórica nominal
Ocupação declarada do cliente	Categórica nominal
Estado Civil declarada do cliente	Categórica nominal
Total de dias na lista de espera do cadastramento	Numérico
Tipo de residência declarada	Categórica nominal
Forma de pagamento principal nas compras	Categórica nominal
Bandeira principal	Categórica nominal
Quantidade de e-mail cadastrado com chave pix	Numérico
Quantidade de telefone cadastrado como chave pix	Numérico
CPF cadastrado como chave pix	Numérico
Quantidade de números aleatórios usados como chave pix	Numérico
Quantidade de chaves pix cadastradas	Numérico
Estado declarado pelo cliente	Categórica nominal
Região do cliente	Categórica nominal
UF declarada versus UF a partir do CPF	Dicotômica
Cliente estava conectado no Wi-fi	Dicotômica
Valor total dos pix transferidos	Numérico

Continua na próxima página

Tabela 5.2 – continuação da página anterior

Descrição	Tipo do Dado
Transferência de pix realizada para uma conta com cpf diferente	Dicotômica
Transferência de pix realizada para uma conta com o mesmo cpf	Dicotômica
Transação realizadas entre pessoas físicas	Dicotômica
Transação realizada de conta jurídica para física	Dicotômica
Transação realizada de conta física para jurídica	Dicotômica
Transação entre contas de pessoas jurídicas	Dicotômica
Chave principal do cliente é e-mail	Dicotômica
Chave principal do cliente é número aleatório	Dicotômica
Chave principal é o cpf	Dicotômica
Chave principal número de telefone	Dicotômica
Total de transferências realizadas	Numérico
Total de bancos cadastrados	Numérico
Valor total que entrou via pix	Numérico
Recebimento de pix de um cpf distinto	Dicotômica
Recebimento de pix do mesmo cpf	Dicotômica
Transação recebida entre pessoas físicas	Dicotômica
Transação recebida de pessoa física para conta jurídica	Dicotômica
Transação recebida de conta jurídica para pessoa física	Dicotômica
Transação recebida entra contas jurídicas	Dicotômica

Para o processamento e entendimento inicial dessas informações, os seguintes critérios foram estabelecidos:

1. **Remoção de variáveis com dados ausentes**: As variáveis que apresentavam mais de 40% de dados ausentes foram removidas da base de dados, resultando na exclusão de 7 variáveis em relação ao total inicial.

2. **Imputar valores faltantes**: Para as variáveis com menos de 40% de dados ausentes, foi imputada a mediana de cada respectiva variável. As variáveis foram analisadas individualmente e apenas uma precisou deste tratamento.

3. **Normalização**: Todas as variáveis numéricas foram normalizadas a partir da função *StandardScaler*.

4. **Categóricas**: Para as variáveis categóricas da base de dados foram criadas variáveis dicotômicas (*dummies*), onde 1 representa a existência do item e 0 a ausência do mesmo.

5. **Correlação**: Foi definido que as variáveis com correlação de *Pearson* superior a 70% deveriam ser removidas da base de modelagem[42].

6. **Faixas**: Tendo em vista que serão testadas diferentes técnicas no momento da modelagem, para algumas variáveis, criou-se uma versão da variável distribuída em faixas.

   A tabela 5.3 a seguir, descreve quais foram as variáveis transformadas em faixas:

Tabela 5.3: Variáveis preditoras em faixa

Item
Faixa do valor total que entrou via pix
Faixa de recebimento de pix de um cpf distinto
Faixa de quantidade de recebimento de pix do mesmo cpf
Faixa de quantidade de transações recebidas entre pessoas físicas
Faixa da quantidade de chaves aleatórias
Faixa da quantidade de pix que saíram da conta
Faixa da quantidade de bancos cadastrados
Faixa de tempo de relacionamento com a instituição
Faixa de idade
Faixa de total de pagamentos do cdc
Faixa do total de pagamentos de boletos
Faixa do total de quantidade de pix que saíram da conta
Faixa do total de cartão físico
Faixa do total de cartão virtual
Faixa do total de pagamentos de conta
Faixa do total de recargas de celular
Faixa do total de pagamento de contas de água
Faixa do total de pagamento de contas telefônicas
Faixa do total de pagamento de tributos governamentais
Faixa do valor total de boletos
Faixa do valor total de compras
Faixa do total de quantidade de compras
Faixa do saldo inicial na conta
Continua na próxima página

Tabela 5.3 – continuação da página anterior

Item
Faixa do máximo de parcelas
Faixa da quantidade de contratos
Faixa do valor total que entrou saiu da conta pix
Faixa de pagamento de pix de um cpf distinto
Faixa de quantidade de pagamento de pix do mesmo cpf
Faixa de quantidade de transações recebidas entre pessoas físicas
Faixa de pagamentos da quantidade de chaves aleatórias
Faixa de quantidade de chaves aleatórias para pagamento
Faixa do total de pagamentos de pix

7. **Seleção de variáveis**: A seleção de variáveis (*feature selection*) é uma etapa importante em muitos modelos de aprendizado de máquina, onde se busca identificar as variáveis mais relevantes para o modelo. No caso deste estudo, foi aplicado a seleção de variáveis via regressão logística com o método *forward*, cujo objetivo é selecionar de forma iterativa as variáveis que melhor explicam a variável dependente do modelo [42]. O *Sequential Feature Selector* da biblioteca *scikit-learn* do *Python* [53] escolhe as variáveis avaliando iterativamente o desempenho do modelo de regressão logística para cada conjunto de variáveis.

Destaca-se que, neste momento, as variáveis relacionadas ao auxílio emergencial ainda não serão incorporadas ao modelo. A Fase I consiste em desenvolver os modelos sem essas variáveis, e somente na Fase II elas serão testadas. No entanto, como o processo de tratamento das variáveis do auxílio emergencial seguiu um procedimento semelhante ao das variáveis originais da base de dados, com exceção dos critérios 5 e 7 mencionados anteriormente, apresentaremos na sequência as variáveis do auxílio emergencial (Tabela 5.4) que serão avaliadas posteriormente neste trabalho.

Tabela 5.4: Variáveis relacionadas ao Auxílio Emergencial que serão testadas na Fase II

Variável	**Descrição**
Tot_Valor_Beneficio	Valor total do benefício recebido
Tot_Pgtos	Quantidade total de pagamentos
Flag_Auxilio	Indicador binário de recebimento do benefício
Tot_Valor_Beneficio_Norm	Valor total do benefício normalizado
Tot_Pgtos_Norm	Quantidade total de pagamentos normalizada
Flag_Auxilio_Norm	Indicador binário de recebimento do benefício normalizado
Fx_Tot_Valor_Beneficio_Menor_986	Valor de benefício menor que R$ 986
Fx_Tot_Valor_Beneficio_Entre_986_E_2513	Valor de benefício entre R$ 986 e R$ 2.513
Fx_Tot_Valor_Beneficio_Entre_2513_E_3658	Valor de benefício entre R$ 2.513 e R$ 3.658
Fx_Tot_Valor_Beneficio_Entre_3658_E_5167	Valor de benefício entre R$ 3.658 e R$ 5.167
Fx_Tot_Valor_Beneficio_Acima_De_5167	Valor de benefício acima de R$ 5.167
Fx_Tot_Pgtos_Menor_2	Faixa de quantidade de pagamentos do benefício menor que 2
Fx_Tot_Pgtos_Entre_2_E_4	Faixa de quantidade de pagamentos do benefício entre 2 e 4
Fx_Tot_Pgtos_Entre_4_E_5ponto5	Faixa de quantidade de pagamentos do benefício entre 4 e 5,5
Fx_Tot_Pgtos_Entre_5pont5_E_7	Faixa de quantidade de pagamentos do benefício entre 5,5 e 7
Fx_Tot_Pgtos_Acima_De_7	Faixa de quantidade de pagamentos do benefício acima de 7
Fx_Tot_Pgtos	Faixa geral de quantidade de pagamentos do benefício
Fx_Tot_Valor_Beneficio	Faixa geral de valor do benefício de auxílio

A imagem a seguir (Figura 5.5), exemplifica todas as etapas mencionadas anteriormente que precedem ao processo de modelagem dos dados:

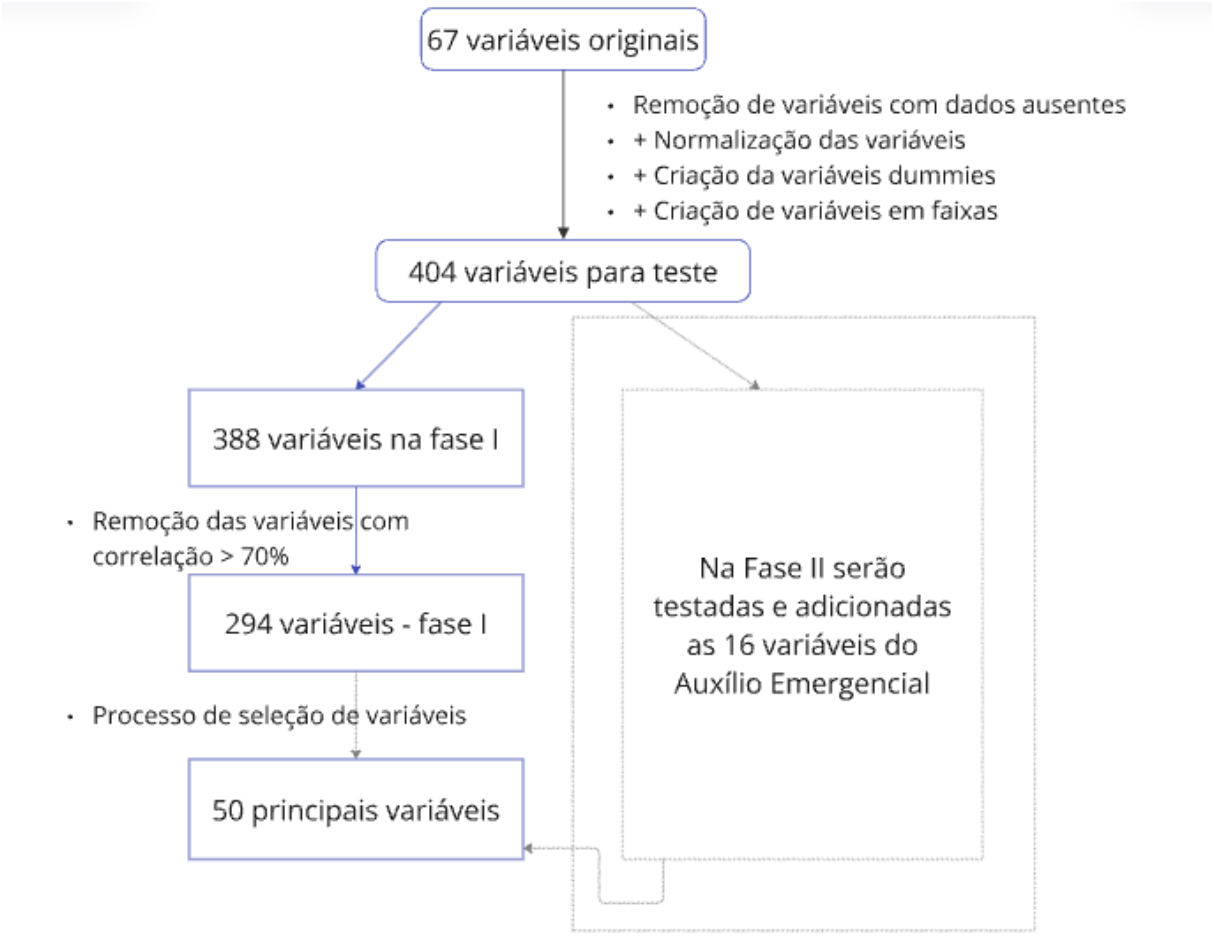


Figura 5.5: Fluxo de tratamento das variáveis.

5.4 Modelagem

A etapa de modelagem é crucial para garantir a construção de modelos preditivos robustos e eficientes, levando em consideração as características da base de dados e os objetivos deste trabalho. Esta fase será dividida em subetapas, que descrevem o processo de separação da base, ajuste de hiperparâmetros, técnicas de balanceamento e avaliação de desempenho, conforme detalhado a seguir.

1. **Separação da Base de Dados**

    - **Divisão Treino/Teste**: a base de dados será dividida em 70% para treino e 30% para teste. A divisão será feita de maneira estratificada, para garantir

que a proporção da variável resposta (que está desbalanceada) seja mantida nas amostras de treino e teste.

2. **Técnicas de Modelagem com ajuste de hiperparâmetros**

   (a) **Regressão Logística**: na regressão logística foram testados os seguintes hiperparâmetros usando o Optuna:

      i. **Penalty**: ['l2', 'l1', None], que define a regularização aplicada ao modelo para evitar sobreajuste, onde **l2** é a Regularização L2 (Ridge), que penaliza grandes coeficientes; **l1** é a Regularização L1 (Lasso), que pode forçar coeficientes a zero, ajudando na seleção de variáveis e **None** é quando não há regularização, o que pode aumentar o risco de sobreajsute [54].

      ii. **Class_Weight**:[None, 'balanced'], que lida com o desbalanceamento das classes, onde **None** é sem ajuste de pesos para as classes e **balanced** ajusta os pesos das classes inversamente proporcionais à sua frequência, ajudando a melhorar a performance em bases desbalanceadas.

      iii. **C**: [0.01, 10], controla a força da regularização, sendo o inverso da regularização (ou seja, quanto maior o valor de C, menor a regularização). Neste projeto foi utilizada a função log uniforme no intervalo 0.01 a 10, que significa que o Optuna tentará valores entre 0.01 e 10 de forma logarítmica para ajustar os melhores parâmetros.

      iv. **Solver**:['lbfgs', 'liblinear', 'newton-cg', 'saga'], que é o algoritmo usado para otimizar a função de custo da Regressão Logística. **lbfgs** é um otimizador eficiente para conjuntos de dados pequenos e médias dimensões; **liblinear** é melhor para pequenas bases; **newton-cg** é um método baseado em gradientes, eficiente para L2 e **saga** é a variante que suporta grandes conjuntos de dados e é eficiente tanto para L1 quanto L2 [55]. A Tabela 5.5 a seguir, sinteitza essas informações:

Tabela 5.5: Comparação dos Solvers na Regressão Logística

Solver	**Características**	**Regularização Suportada**	**Aplicabilidade**
`lbfgs`	Otimizador eficiente para conjuntos de dados de pequenas e médias dimensões.	L2	Recomendado para problemas multiclasse e conjuntos de dados de tamanho moderado.
`liblinear`	Baseado no método de descida do gradiente coordenado, é adequado para conjuntos de dados pequenos e suporta regularização L1 e L2.	L1 e L2	Ideal para problemas de classificação binária e quando a seleção de variáveis é necessária devido à regularização L1.
`newton-cg`	Método baseado em gradientes que utiliza a matriz Hessiana completa, proporcionando convergência precisa.	L2	Adequado para problemas multiclasse e conjuntos de dados de tamanho moderado, onde a precisão é crucial.
`saga`	Variante do método SAG que suporta grandes conjuntos de dados e é eficiente tanto para regularização L1 quanto L2.	L1 e L2	Recomendado para grandes conjuntos de dados e problemas multiclasse.

v. **Max_iter**: [20, 150], define o número máximo de iterações que o algoritmo de otimização vai executar. Neste trabalho, o número escolhido foi baseado em modelos anteriores da instituição e deverá estar dentro do intervalo entre 20 e 150. Um número muito baixo pode fazer com que o algoritmo não convirja, enquanto um número muito alto pode aumentar o tempo de execução desnecessariamente.

(b) **Floresta Aleatória**: na floresta aleatória foram testados os seguintes hiperparâmetros:

i. **N_estimators**: [50, 200], define o número de árvores na floresta. Um valor maior geralmente melhora a performance do modelo, mas também aumenta o tempo de execução. Neste trabalho, o valor escolhido foi baseado em modelos anteriores da instituição e o número de árvores será escolhido no intervalo entre 50 e 200.

ii. **Max_depth**: [5, 15], controla a profundidade máxima de cada árvore. Profundidades maiores permitem que o modelo aprenda mais detalhes dos dados, mas também podem levar ao sobreajuste. Neste projeto, baseado em modelos anteriores, o valor ideal de profundidade será testado entre 5 e 15, com profundidades menores generalizando melhor, enquanto profundidades maiores permitem modelos mais complexos.

iii. **Min_samples_split**: [2, 20], determina o número mínimo de amostras necessárias para dividir um nó. Um valor menor pode levar a um modelo mais detalhado, mas pode aumentar o risco de sobreajuste. Um valor maior faz o modelo ser mais conservador, o que pode reduzir a chance de sobreajuste, mas pode não capturar toda a complexidade dos dados. Neste trabalho, baseado em modelos anteriores, o número mínimo de amostras para divisão será ajustado entre 2 e 20.

iv. **Min_samples_leaf**: [1, 5], especifica o número mínimo de amostras que uma folha deve ter. Um valor baixo pode permitir folhas muito específicas (com alto risco de sobreajuste), enquanto valores mais altos tornam as folhas mais generalizadas. Aqui, o número mínimo de amostras por folha será ajustado entre 1 e 5.

(c) **LightGBM**: para o LightGBM foram testados os seguintes hiperparâmetros:

i. **Num_leaves**: [10, 30], define o número máximo de folhas em cada árvore. Um número maior de folhas aumenta a complexidade do modelo, permitindo que ele aprenda mais padrões, mas também aumenta o risco de sobreajuste. O intervalo de 10 a 30 foi definido a partir de modelos anteriores para encontrar o equilíbrio entre complexidade e generalização.

ii. **Learning_rate**: [0.01, 0.1, 0.5], controla o quanto o modelo ajusta os pesos de cada iteração durante o processo de aprendizado. Um valor menor de **Learning_rate**, como 0.01, faz o modelo aprender de forma mais lenta e cuidadosa, enquanto valores maiores, como 0.5, aceleram o aprendizado, mas podem levar ao sobreajuste. Neste projeto, baseado em modelos anteriores, testaremos os valores 0.01, 0.1 e 0.5.

iii. **N_estimators**: [100, 300], define o número de árvores a serem construídas. Valores maiores tendem a melhorar a performance do modelo, porém

aumentam o tempo de execução. O intervalo de 100 a 300 será utilizado para balancear a qualidade da predição e o tempo de treinamento. Neste trabalho,o número escolhido foi baseado em modelos anteriores da instituição.

(d) **AdaBoost com custo**: no AdaBoost foram testados os seguintes hiperparâmetros:

i. **N_estimators**: [50, 300], que define o número de estimadores (árvores fracas) a serem construídos. Um valor maior de estimadores pode aumentar a complexidade do modelo e melhorar a performance, mas também aumenta o tempo de treinamento. O intervalo de 50 a 300 foi escolhido, baseado em experimentos anteriores, para buscar um equilíbrio entre performance e eficiência.

ii. **Learning_rate**: [0.01, 1.0], controla o peso de cada estimador no resultado final. Valores menores de **learning_rate** tornam o modelo mais conservador, enquanto valores maiores aceleram o aprendizado, mas podem aumentar o risco de sobreajuste. Esse intervalo permite avaliar diferentes níveis de aprendizado.

iii. **Max_depth**: [1, 10], define a profundidade máxima das árvores individuais. Árvores mais profundas podem capturar mais padrões, mas também correm o risco de sobreajuste. Considerando experimentos realizados anteriormente na instituição, testaremos entre profundidades de 1 a 10 para encontrar o equilíbrio ideal entre simplicidade e complexidade.

Além dos hiperparâmetros tradicionais, serão aplicadas duas diferentes **matrizes de custo** para considerar o impacto de falsos negativos mais custosos, ajustando assim o modelo para lidar com o desbalanceamento da base de dados de maneira mais eficiente. Neste estudo foram considerados dois cenários para as matrizes custo, definidos a partir do alinhamento internos com os especialistas da instituição. Uma em que os falsos negativos serão penalizados em 5x, e outra em que eles serão penalizados em 10x em relação aos falsos positivos. As matrizes de custo são aplicadas da seguinte forma:

i. **Matriz de Custo 1**: falsos negativos são 5 vezes mais custosos que falsos positivos. A matriz de custo é definida como:

$$\text{Matriz de custo 1} = \begin{bmatrix} 0 & 1 \\ 5 & 0 \end{bmatrix}$$

Com essa matriz, o Optuna ajustará os parâmetros do modelo de forma a minimizar os falsos negativos, considerando que eles têm um impacto maior.

ii. **Matriz de Custo 2**: Falsos negativos são 10 vezes mais custosos que falsos positivos. A matriz de custo é definida como:

$$\text{Matriz de custo 2} = \begin{bmatrix} 0 & 1 \\ 10 & 0 \end{bmatrix}$$

Essa matriz de custo impõe um peso ainda maior sobre os falsos negativos, forçando o modelo a focar ainda mais em sua redução durante o ajuste de hiperparâmetros.

(e) **Naive Bayes com custo**: para o Naive Bayes, foi implementada uma abordagem de ajuste sensível a custos. Isso significa que, além de ajustar os parâmetros, foi levado em consideração o impacto de falsos negativos e falsos positivos, utilizando diferentes matrizes de custo. O objetivo é minimizar o impacto de falsos negativos, especialmente em cenários onde eles são mais prejudiciais.

i. **Função objetiva**: a função objetiva para o Optuna foi configurada para maximizar a métrica de AUC (Área Sob a Curva) com base na predição das probabilidades de cada classe. A função utiliza o Naive Bayes sensível a custos, treinando o modelo e retornando o AUC para otimização.

ii. **Matriz de Custo 1**: Neste cenário, os falsos negativos são 5 vezes mais custosos que os falsos positivos. A matriz de custo é definida como:

$$\text{Matriz de custo 1} = \begin{bmatrix} 0 & 1 \\ 5 & 0 \end{bmatrix}$$

Isso garante que o modelo dê maior importância a reduzir os falsos negativos durante o ajuste dos parâmetros.

iii. **Matriz de Custo 2**: Neste segundo cenário, os falsos negativos são 10 vezes mais custosos que os falsos positivos, forçando o modelo a priorizar ainda mais a redução de falsos negativos. A matriz de custo é definida como:

$$\text{Matriz de custo 2} = \begin{bmatrix} 0 & 1 \\ 10 & 0 \end{bmatrix}$$

Em ambos os cenários, foi utilizado o Optuna para otimizar o modelo Naive Bayes sensível a custos, com 50 tentativas de ajuste. O melhor conjunto de parâmetros foi obtido para cada matriz de custo, garantindo que o modelo leve

em consideração o impacto dos falsos negativos no modelo.

3. **Cenários de Balanceamento**: Conforme explicado no capítulo 4, serão testados três cenários de balanceamento dos dados:
   (a) **Sem ajuste de desbalanceamento**: o modelo será treinado com a base original.
   (b) **SMOTE**: será aplicado o sobreamostragem sintético para equilibrar as classes minoritárias da variável resposta.
   (c) **Subamostragem**: será aplicada a técnica de redução da classe majoritária para equilibrar a variável resposta na base de dados.
4. **Métricas e técnica de Avaliação** Para todos os cenários testados, serão calculadas as seguintes métricas nas bases de treino, teste e K-fold:
   (a) **Acurácia**: proporção de previsões corretas entre o total de previsões.
   (b) **Precisão**: proporção de verdadeiros positivos entre as previsões positivas.
   (c) **Sensibilidade**: proporção de verdadeiros positivos em relação a todas as instâncias verdadeiramente positivas
   (d) **F1-score**: média harmônica entre precisão e sensibilidade, balanceando ambas as métricas.
   (e) **KS**: distância máxima entre as distribuições acumuladas de classes.
   (f) **Especificidade**: proporção de negativos corretos entre todos os que são realmente negativos.
   (g) **G-mean**: raiz quadrada do produto entre sensibilidade e especificidade, medindo o equilíbrio entre elas.
   (h) **AUC**: área sob a curva ROC, medindo a separação entre classes.
   (i) **Kappa**: avaliação da concordância entre previsões e rótulos reais ajustada pelo acaso.
   (j) **K-fold**: Para cada combinação de técnica de aprendizado de máquinas e cenários de ajustes de desbalanceamento, foi calculado o K-fold, como uma validação cruzada da técnica testada.

Após as etapas descritas anteriormente (Figura 5.5), 50 variáveis explicativas foram selecionadas para serem testadas nos modelos de aprendizado de máquina. No entanto, durante a validação dos modelos, foi identificado um possível problema de multicolinearidade, o que poderia impactar negativamente a estabilidade e interpretação dos modelos.

Para verificar e quantificar esse problema, utilizou-se o Fator de Inflação da Variância (VIF).

O VIF é uma medida estatística que avalia o grau de colinearidade entre as variáveis explicativas. Ele indica o quanto a variância de um coeficiente é inflacionada devido à correlação com outras variáveis. Valores de VIF superiores a 10, por exemplo, sugerem alta multicolinearidade, o que pode dificultar a interpretação precisa dos coeficientes do modelo [42]. Durante a análise, uma variável (Faixa do total de pagamentos de conta) apresentou um VIF superior a 18, sendo, portanto, excluída para evitar efeitos adversos na modelagem.

Posteriormente, foi aplicada a técnica de Regressão Logística com Eliminação Recursiva de Características (RFE) nas 49 variáveis restantes, considerando dois cenários distintos: um com a seleção de 25 variáveis e outro com 30 variáveis. A RFE é um método de seleção de variáveis em que o modelo é treinado repetidamente, removendo, em cada iteração, as variáveis menos importantes com base em sua contribuição para o desempenho do modelo. O objetivo da RFE é identificar o subconjunto de variáveis mais relevantes, reduzindo a complexidade do modelo sem comprometer sua performance preditiva [56][57].

Para definir se seriam as 25 ou 30 variáveis selecionadas, o processo de modelagem foi realizado até o final, onde é possível gerar o gráfico de importância das variáveis. Com isso, identificou-se que as 5 variáveis adicionais do conjunto de 30, ficaram na parte inferior do gráfico, trazendo pouca importância para o modelo. Logo, seguimos o processo com as 25 variáveis selecionadas pela RFE.

Após o RFE foi aplicado novamente o VIF para checar a multicolinearidade, eliminando mais duas variáveis (Saldo inicial de até R$300,00 e Quantidade de chaves Pix cadastradas pelo CPF). Com isso, a Fase I foi finalizada com a seleção de 23 variáveis explicativas, consideradas as mais adequadas para os modelos, conforme mostra a Tabela 5.6 a seguir:

Tabela 5.6: Variáveis nos modelos da Fase I

Variável
ESTADO_CIVIL_TRAT_CASADO
FX_SLD_INICIAL_TRAT_De600a1000
CHAVE_NATIONALREGISTRATION
QTD_CAD_CHAVE_PIX_TELEFONE
Continua na próxima página

Tabela 5.6 – continuação da página anterior

Variável
FX_TOT_VLR_BOLETO_TRAT_menorigual_400
IDADE
FX_TOT_PAG_CONTA_GOVERNO_TRAT_0
FX_MAX_PARCELAS_TRAT_De15a20parcelas
FX_QTDE_BANCO_PIXIN_TRAT_De3a5
FX_TRANSACAO_P2P_PIXOUT_TRAT_Acima10P2P
REGIAO_TRAT_Norte
OCUPACAO_TRAT_COOPERADO
REGIAO_TRAT_Nordeste
DIRETORIA_TRAT_DIR4
FX_AMOUNT_PIXOUT_TRAT_De3000a5000
FX_SLD_INICIAL_TRAT_De3500a5000
OCUPACAO_TRAT_AUTONOMO
FX_MAX_PARCELAS_TRAT_Ate3parcelas
FX_SLD_INICIAL_TRAT_De300a600
UF_CPF_TRAT_3
FX_TOT_ENTRADAS_TRAT_menorigual_400
ESTADO_CIVIL_TRAT_DIVORCIADO
ESTADO_CIVIL_TRAT_SOLTEIRO

Para Fase II, como mencionado anteriormente foram incluídas 16 variáveis do auxílio emergencial. Além dos tratamentos mencionados nos tópicos anteriores, as variáveis do auxílio emergencial foram selecionadas a partir do uso da técnica de CatBoost.

Após a aplicação desta técnica as variáveis selecionadas do auxílio emergencial e que serão testadas no modelo são descritas na Tabela 5.7:

Tabela 5.7: Variáveis do Auxílio para Fase II

Variável
FLAG_AUXILIO_NORM
FX_TOT_VALOR_BENEFICIO_menor_986
FX_TOT_VALOR_BENEFICIO_entre_986_e_2513
Continua na próxima página

Tabela 5.7 – continuação da página anterior

Variável
FX_TOT_VALOR_BENEFICIO_entre_2513_e_3658
FX_TOT_VALOR_BENEFICIO_entre_3658_e_5167
FX_TOT_VALOR_BENEFICIO_acima_de_5167
FX_TOT_PGTOS_menor_2
FX_TOT_PGTOS_entre_2_e_4
FX_TOT_PGTOS_entre_4_e_5ponto5

5.5 Resultados

Nesta seção serão apresentados os resultados dos modelos desenvolvidos na Fase I, bem como, o resultado da incorporação dos dados Auxílio Emergencial no modelo escolhido.

A partir da Tabela 5.8 pode-se avaliar que a regressão logística apresentou, de maneira geral, um desempenho inconstante em relação as demais técnicas principalmente no que tange a sensibilidade e o KS do modelo. Em todos os cenários (Sem correção, *SMOTE* e Subamostragem), a regressão logística não conseguiu capturar bem a classe minoritária, haja vista um valor de sensibilidade quase nulo nos cenários sem técnicas de correção de desbalanceamento. Outro fator que pesou negativamente na regressão logística foi um valor de especificidade igual a 1 em todos os cenários, significando que o modelo estava altamente enviesado para prever a classe majoritária (clientes adimplentes), resultando em um alto valor de acurácia para esta classe.

Por sua vez, as combinações de modelos a partir das florestas aleatórias tiveram resultados mais promissores, mesmo mostrando uma grande variação dos resultados dependendo da técnica de correção de desbalanceamento utilizada. Os melhores resultados foram obtidos com *SMOTE* e subamostragem. A aplicação do *SMOTE* fez com que o resultado da árvore melhorasse o patamar da sensibilidade, saindo de 0.0 no cenário sem correção para 0.399 quando olhamos a base de teste. Outro ponto que indica que o balanceamento foi capaz de capturar melhor os clientes inadimplentes é a variação da sensibilidade entre as bases de treino, teste e o k-fold. O KS também aumentou em comparação com o cenário sem balanceamento, mas ainda com uma queda no teste, sugerindo que o modelo pode ter se ajustado melhor aos dados de treino do que ao teste. Além disso, os valores

de especificidade e G-mean também melhoraram em relação ao cenário sem correção do desbalanceamento da variável resposta.

Ao analisar os resultados obtidos pelo *LightGBM*, nota-se que seus resultados foram superiores em relação a Regressão Logística e alternados em relação as florestas aleatórias. O *LightGBM* no *SMOTE* teve um valor de sensibilidade melhor do que os resultados das florestas aleatórias, mas também com oscilações em relação as bases de treino e teste. Na correção usando a técnica de subamostragem, o *LightGBM* tem um desempenho mais estável, mantendo bons patamares de sensibilidade, KS, Acurácia, G-mean nas bases treino, teste e também no k-fold. Contudo, ao comparamos a sensibilidade do *LightGBM* e florestas aleatórias na subamostragem, os resultados do *LightGBM* foram inferiores.

Tabela 5.8: Resultados dos modelos sem custo

Modelo	Base	Acurácia	Precisão	Sensibilidade	F1-Score	KS	Especificidade	G-mean	AUC-ROC	Kappa
Sem correção										
Logística	Treino	0.771	0.500	0.000	0.000	0.092	1.000	0.000	0.563	0.000
	Teste	0.771	0.286	0.000	0.000	0.096	1.000	0.000	0.565	0.000
	K-fold	0.771	0.543	0.000	0.000	0.059	1.000	0.000	0.563	0.000
Florestas Aleatórias	Treino	0.771	0.972	0.001	0.002	0.110	1.000	0.001	0.579	0.002
	Teste	0.771	0.286	0.000	0.000	0.097	1.000	0.000	0.567	0.000
	K-fold	0.771	0.963	0.001	0.002	0.059	1.000	0.001	0.579	0.002
LightGbm	Treino	0.771	0.841	0.002	0.004	0.112	1.000	0.002	0.580	0.003
	Teste	0.771	0.403	0.001	0.001	0.096	1.000	0.001	0.566	0.001
	K-fold	0.771	0.833	0.002	0.004	0.059	1.000	0.002	0.582	0.003
SMOTE										
Logística	Treino	0.546	0.548	0.529	0.538	0.092	0.564	0.298	0.567	0.092
	Teste	0.554	0.262	0.521	0.349	0.087	0.564	0.294	0.557	0.064
	K-fold	0.546	0.548	0.530	0.538	0.067	0.563	0.298	0.567	0.092
Florestas Aleatórias	Treino	0.558	0.579	0.427	0.492	0.119	0.690	0.295	0.589	0.117
	Teste	0.622	0.276	0.399	0.326	0.088	0.689	0.275	0.557	0.076
	K-fold	0.560	0.572	0.475	0.519	0.067	0.644	0.305	0.590	0.119
LightGbm	Treino	0.563	0.574	0.492	0.529	0.127	0.635	0.312	0.597	0.126
	Teste	0.592	0.269	0.455	0.338	0.088	0.633	0.288	0.556	0.071
	K-fold	0.563	0.574	0.489	0.528	0.067	0.638	0.312	0.597	0.127
Subamostragem										
Logística	Treino	0.545	0.549	0.506	0.527	0.093	0.585	0.296	0.562	0.091
	Teste	0.568	0.268	0.510	0.351	0.096	0.585	0.298	0.565	0.072
	K-fold	0.545	0.549	0.506	0.527	0.060	0.585	0.296	0.562	0.091
Florestas Aleatórias	Treino	0.536	0.529	0.652	0.584	0.073	0.420	0.274	0.554	0.072
	Teste	0.467	0.247	0.646	0.357	0.052	0.419	0.266	0.536	0.031
	K-fold	0.536	0.527	0.648	0.582	0.061	0.421	0.274	0.554	0.073
LightGbm	Treino	0.543	0.546	0.503	0.523	0.099	0.588	0.297	0.570	0.092
	Teste	0.570	0.269	0.518	0.355	0.088	0.586	0.295	0.564	0.072
	K-fold	0.543	0.545	0.503	0.523	0.060	0.588	0.297	0.570	0.092

Como mencionado nos capítulos anteriores, a utilização de técnicas com custos foi testada porque, em cenários de modelagem, há a necessidade de balancear as classes quando ocorre um desbalanceamento da variável resposta. Essa prática é importante principalmente em problemas de classificação onde uma das classes dominam, resultando em um

viés dos modelos para essas categorias, comprometendo a sensibilidade e a especificidade para a classe minoritária.

Ao analisar os resultados da Tabela 5.9, onde foram consideradas as matrizes de custo, observa-se que o *Naive Bayes* no cenário da matriz de custo 5x apresentou resultados consistentes entre treino, teste e K-fold, com uma precisão moderada (Acurácia: 0.738, 0.679, 0.677) e especificidade com valores de 0.931, 0.811 e 0.805, respectivamente. Contudo, nas demais métricas de avaliação o modelo apesar de estabilidade entre as bases apresentou um patamar de resultados inferior ao esperado. Nos cenários com *SMOTE* e subamostragem, os resultados do *Naive Bayes* na matriz de custo 5x não foram promissores.

Ao passar a matriz de custo do *Naive Bayes* para penalizar em 10x, onde os falsos negativos são 10 vezes mais custosos que falsos positivos, a performance do *Naive Bayes* degrada significativamente entre as bases de treino e teste, indicando que o modelo não está atingindo a capacidade preditiva esperada.

A segunda técnica com custo avaliada foi o *AdaBoost*, também nas matrizes de custo 5x e 10x. Neste caso, o modelo apresentou uma performance em patamares inferiores em relação as demais técnicas testadas, mesmo demonstrando uma estabilidade entre as bases de treino, teste e validação. Ao analisar o *AdaBoost* nos cenários *SMOTE* e subamostragem, os resultados foram instáveis entre as bases de treino, teste e k-fold, em todas as métricas avaliadas. Além disso, houve pouca distinção dos resultados obtidos considerando a correção da variável resposta com *SMOTE* e *AdaBoost*.

Em linhas gerais, nos cenários analisados, os resultados com ou sem técnicas de correção, o *Naive Bayes* não consegue performar bem em termos de precisão e *F1-score*, mas melhora marginalmente com *SMOTE* e subamostragem. Tais correções da variável resposta aumentam a sensibilidade, mas o modelo continua tendo dificuldade em lidar com desequilíbrios significativos. Por sua vez, o *AdaBoost*, mesmo com técnicas de correção como *SMOTE* e subamostragem, tem um desempenho consistente e baixo, com pouca variação nas métricas. O modelo tende a classificar quase todas as observações como pertencentes à classe majoritária, com alta sensibilidade, mas baixa precisão e acurácia.

Tabela 5.9: Resultados dos modelos com custo

Modelo	Base	Acurácia	Precisão	Sensibilidade	F1-Score	KS	Especificidade	G-mean	AUC-ROC	Kappa
Sem correção										
Naive Bayes 5x	Treino	0.738	0.283	0.092	0.139	0.086	0.931	0.086	0.551	0.030
	Teste	0.679	0.272	0.237	0.253	0.090	0.811	0.192	0.551	0.050
	K-fold	0.677	0.271	0.244	0.257	0.088	0.805	0.196	0.551	0.051
Naive Bayes 10x	Treino	0.738	0.283	0.092	0.139	0.086	0.931	0.086	0.551	0.030
	Teste	0.486	0.252	0.632	0.360	0.090	0.443	0.280	0.551	0.049
	K-fold	0.449	0.248	0.689	0.364	0.088	0.378	0.258	0.551	0.042
AdaBoost 5x	Treino	0.242	0.231	0.987	0.374	0.095	0.020	0.020	0.565	0.004
	Teste	0.229	0.229	1.000	0.373	0.098	0.000	0.000	0.567	0.000
	K-fold	0.229	0.229	1.000	0.373	0.096	0.000	0.000	0.565	0.000
AdaBoost 10x	Treino	0.229	0.229	1.000	0.373	0.095	0.000	0.000	0.566	0.000
	Teste	0.229	0.229	1.000	0.373	0.099	0.000	0.000	0.567	0.000
	K-fold	0.229	0.229	1.000	0.373	0.096	0.000	0.000	0.565	0.000
SMOTE										
Naive Bayes 5x	Treino	0.521	0.582	0.147	0.235	0.089	0.894	0.132	0.559	0.041
	Teste	0.658	0.267	0.282	0.274	0.080	0.770	0.217	0.548	0.050
	K-fold	0.534	0.564	0.299	0.391	0.090	0.768	0.230	0.559	0.068
Naive Bayes 10x	Treino	0.521	0.582	0.147	0.235	0.089	0.894	0.132	0.559	0.041
	Teste	0.492	0.253	0.621	0.359	0.080	0.454	0.282	0.548	0.049
	K-fold	0.542	0.536	0.631	0.579	0.090	0.452	0.284	0.559	0.083
AdaBoost 5x	Treino	0.500	0.500	1.000	0.667	0.145	0.000	0.000	0.604	0.000
	Teste	0.229	0.229	1.000	0.373	0.094	0.000	0.000	0.559	0.000
	K-fold	0.229	0.229	1.000	0.373	0.095	0.000	0.000	0.564	0.000
AdaBoost 10x	Treino	0.500	0.500	1.000	0.667	0.145	0.000	0.000	0.604	0.000
	Teste	0.229	0.229	1.000	0.373	0.094	0.000	0.000	0.559	0.000
	K-fold	0.229	0.229	1.000	0.373	0.095	0.000	0.000	0.564	0.000
Subamostragem										
Naive Bayes 5x	Treino	0.522	0.561	0.204	0.300	0.086	0.840	0.172	0.551	0.044
	Teste	0.298	0.233	0.898	0.370	0.090	0.119	0.107	0.552	0.009
	K-fold	0.509	0.505	0.894	0.646	0.087	0.124	0.110	0.549	0.018
Naive Bayes 10x	Treino	0.522	0.561	0.204	0.300	0.086	0.840	0.172	0.551	0.044
	Teste	0.257	0.231	0.960	0.372	0.090	0.049	0.047	0.552	0.004
	K-fold	0.504	0.502	0.956	0.658	0.087	0.052	0.049	0.549	0.007
AdaBoost 5x	Treino	0.500	0.500	1.000	0.667	0.145	0.000	0.000	0.604	0.000
	Teste	0.229	0.229	1.000	0.373	0.095	0.000	0.000	0.559	0.000
	K-fold	0.229	0.229	1.000	0.373	0.095	0.000	0.000	0.564	0.000
AdaBoost 10x	Treino	0.500	0.500	1.000	0.667	0.145	0.000	0.000	0.604	0.000
	Teste	0.229	0.229	1.000	0.373	0.095	0.000	0.000	0.559	0.000
	K-fold	0.229	0.229	1.000	0.373	0.095	0.000	0.000	0.564	0.000

A partir da análise realizada, entende-se que em um primeiro momento os resultados dos modelos com correção da variável resposta com *SMOTE* e subamostragem foram superiores aos resultados sem correção da variável resposta. Para auxiliar na escolha de qual modelo seguir para incorporar as variáveis do auxílio emergencial, nestes dois cenários de ajuste da variável resposta foi calculado o Kappa, conforme ilustra a Figura 5.6 a seguir:

Kappa entre os modelos - SMOTE	Kappa
Logística x Florestas aleatórias	0.646
Logística x LightGBM	0.770
Florestas aleatórias x LightGBM	0.786

Kappa entre os modelos - Subamostragem	Kappa
Logística x Florestas aleatórias	0.758
Logística x LightGBM	0.784
Florestas aleatórias x LightGBM	0.834

Figura 5.6: Kappa entre as técnicas sem custo.

Considerando o Kappa entre as técnicas com custo e suas respectivas matrizes, os resultados foram inferiores aos alcançados com os modelos sem custo. A Figura 5.7 apresenta o Kappa em todos os cenários de correção do desbalanceamento da variável resposta para as técnicas com custo.

Kappa entre os modelos - Sem correção	Kappa
Naive Bayes 5x vs Naive Bayes 10x	1.00
Naive Bayes 5x vs AdaBoost 5x	0.00
Naive Bayes 5x vs AdaBoost 10x	0.00
Naive Bayes 10x vs AdaBoost 5x	0.00
Naive Bayes 10x vs AdaBoost 10x	0.00
AdaBoost 5x vs AdaBoost 10x	0.01

Kappa entre os modelos - SMOTE	Kappa
Naive Bayes 5x vs Naive Bayes 10x	1.00
Naive Bayes 5x vs AdaBoost 5x	0.00
Naive Bayes 5x vs AdaBoost 10x	0.00
Naive Bayes 10x vs AdaBoost 5x	0.00
Naive Bayes 10x vs AdaBoost 10x	0.00
AdaBoost 5x vs AdaBoost 10x	nan

Kappa entre os modelos - SMOTE	Kappa
Naive Bayes 5x vs Naive Bayes 10x	1.00
Naive Bayes 5x vs AdaBoost 5x	0.00
Naive Bayes 5x vs AdaBoost 10x	0.00
Naive Bayes 10x vs AdaBoost 5x	0.00
Naive Bayes 10x vs AdaBoost 10x	0.00
AdaBoost 5x vs AdaBoost 10x	0.00

Figura 5.7: Kappa entre as técnicas com custo.

Com base nas análises anteriores, os modelos mais promissores foram aqueles desenvolvidos utilizando as técnicas de Florestas Aleatórias e *LightGBM*, particularmente nos cenários de *SMOTE* e subamostragem. Diante desse desempenho, uma análise mais aprofundada foi conduzida para avaliar detalhadamente a performance de cada um desses modelos, com o objetivo de identificar qual deles apresenta a melhor adequação para a aplicação da Fase II.

O primeiro gráfico da Figura 5.8 resume o percentual de inadimplência (Grupos e Percentual de 1 - Teste) nos decis da base de teste do modelo de Florestas Aleatórias. A intenção é obter uma ordenação decrescente, de forma que o primeiro decil concentre o maior percentual de inadimplência e que esse percentual vá diminuindo à medida que se

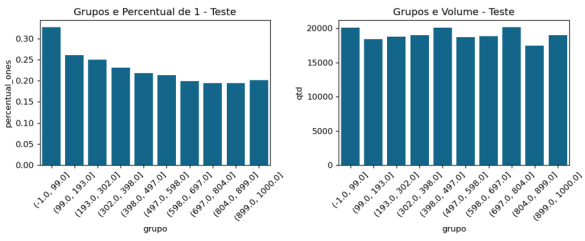
avança para os demais decis. No segundo gráfico (Grupos e Volume - Teste), espera-se uma distribuição uniforme entre os decis. No entanto, alguns desvios são observados, o que descaracteriza a uniformidade desejada.

A tabela da Figura 5.8 detalha a inadimplência em cada decil das Florestas Aleatórias no SMOTE. Para o cálculo da inadimplência, os clientes foram ordenados por sua respectiva probabilidade de inadimplência, depois foi verificado quais eram efetivamente inadimplentes (qtde 1). Em seguida, foi realizado o percentual tomando por base a quantidade efetiva de inadimplentes e a quantidade total em cada decil.

O primeiro decil apresentou uma inadimplência de 32,6%, e o objetivo é que essa inadimplência vá diminuindo ao longo dos decis. Embora tenha ocorrido uma redução até o oitavo decil, nos dois últimos decis essa tendência não se manteve, sugerindo uma possível perda de capacidade preditiva do modelo em identificar os melhores clientes.

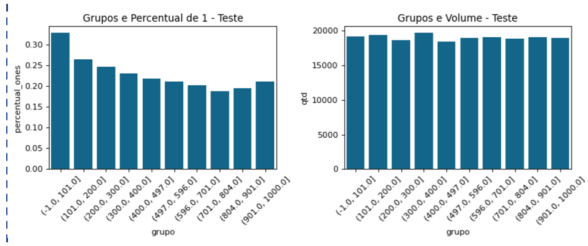
Ao realizar a mesma análise no modelo utilizando o LightGBM, observa-se uma ordenação semelhante no percentual de inadimplência (Grupos e Percentual de 1 - Teste) na base de teste entre os decis, com uma distribuição de volume (Grupos e Volume - Teste) que se aproxima mais de uma distribuição uniforme. Na tabela de inadimplência, também houve uma redução consistente até o oitavo decil. No entanto, ocorreu uma inversão no nono decil, onde a inadimplência subiu de 18,7% para 19,4%. O último decil apresenta uma inadimplência comparável à dos clientes do sexto decil, o que sugere, novamente, uma possível perda de capacidade preditiva do modelo nos melhores clientes.

Smote – Florestas aleatórias



Decil	Intervalo Decil	qtde	qtde 1	percentual de 1
1.0	(-1.0, 99.0]	20107	6563	32,6%
2.0	(99.0, 193.0]	18366	4784	26,0%
3.0	(193.0, 302.0]	18750	4687	25,0%
4.0	(302.0, 398.0]	18970	4384	23,1%
5.0	(398.0, 497.0]	20063	4358	21,7%
6.0	(497.0, 598.0]	18716	3992	21,3%
7.0	(598.0, 697.0]	18805	3750	19,9%
8.0	(697.0, 804.0]	20134	3899	19,4%
9.0	(804.0, 899.0]	17466	3388	19,4%
10.0	(899.0, 1000.0]	18976	3822	20,1%

Smote – LightGBM



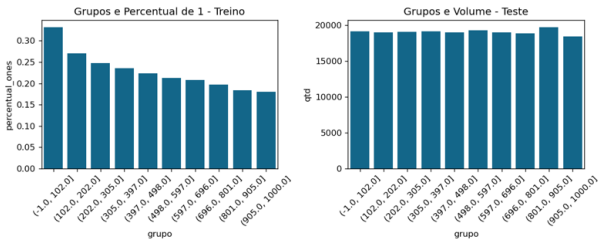
Decil	Intervalo Decil	qtde	qtde 1	percentual de 1
1.0	(-1.0, 101.0]	19184	6280	32,7%
2.0	(101.0, 200.0]	19408	5122	26,4%
3.0	(200.0, 300.0]	18650	4593	24,6%
4.0	(300.0, 400.0]	19674	4531	23,0%
5.0	(400.0, 497.0]	18460	4023	21,8%
6.0	(497.0, 596.0]	18969	3990	21,0%
7.0	(596.0, 701.0]	19108	3860	20,2%
8.0	(701.0, 804.0]	18845	3528	18,7%
9.0	(804.0, 901.0]	19134	3711	19,4%
10.0	(901.0, 1000.0]	18921	3989	21,1%

Figura 5.8: Ordenação em decis considerando o SMOTE.

A Figura 5.9 apresenta os mesmos gráficos e tabelas da figura anterior, agora considerando a correção da variável resposta por meio da subamostragem. No cenário com Florestas Aleatórias, observa-se uma melhora clara na ordenação do modelo. No primeiro gráfico, houve uma ordenação consistente no percentual de inadimplência (Grupos e Percentual de 1 - Teste) entre os grupos de decis. No segundo gráfico (Grupos e Volume - Teste), houve uma melhora significativa na distribuição de volume, resultando em uma quase distribuição uniforme entre os decis. Essa melhora também é evidente na tabela de inadimplência, onde o primeiro decil apresenta uma taxa de inadimplência de 33,1% e o último, 18,1%, sem inversões na ordem de inadimplência ao longo dos decis.

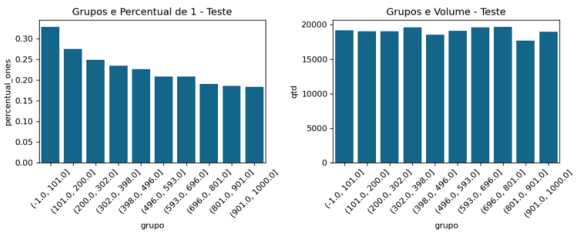
Realizando a mesma análise com a técnica de LightGBM, considerando a correção do desbalanceamento da variável resposta por meio da subamostragem, observa-se uma leve perda na uniformidade da distribuição dos decis em comparação ao cenário com SMOTE. No entanto, houve um ganho na ordenação do percentual de inadimplência (Grupos e Percentual de 1 - Teste) na base de teste, assim como uma melhora na distribuição da inadimplência entre os decis. A inadimplência apresentou uma queda constante, partindo de 32,8% no primeiro decil e chegando a 19,0% no décimo decil, sem inversões.

Subamostragem – Florestas aleatórias



Decil	Intervalo Decil	qtde	qtde 1	percentual de 1
1.0	(-1.0, 102.0]	19227	6355	33,1%
2.0	(102.0, 202.0]	19060	5020	26,3%
3.0	(202.0, 305.0]	19146	4614	24,1%
4.0	(305.0, 397.0]	19165	4406	23,0%
5.0	(397.0, 498.0]	18999	4221	22,2%
6.0	(498.0, 597.0]	19298	4070	21,1%
7.0	(597.0, 696.0]	18949	3931	20,7%
8.0	(696.0, 801.0]	18689	3836	20,5%
9.0	(801.0, 905.0]	19779	3905	19,7%
10.0	(905.0, 1000.0]	18041	3269	18,1%

Subamostragem – LightGBM



Decil	Intervalo Decil	qtde	qtde 1	percentual de 1
1.0	(-1.0, 101.0]	19138	6268	32,8%
2.0	(101.0, 200.0]	19025	5209	27,4%
3.0	(200.0, 302.0]	18996	4715	24,8%
4.0	(302.0, 398.0]	19657	4591	23,4%
5.0	(398.0, 496.0]	18538	4193	22,6%
6.0	(496.0, 593.0]	19081	3965	20,8%
9.0	(801.0, 901.0]	17967	3507	19,5%
10.0	(901.0, 1000.0]	19998	3791	19,0%

Figura 5.9: Ordenação em decis considerando a Subamostragem.

Considerando todos os cenários descritos anteriormente, conclui-se que o modelo mais promissor foi o de Florestas Aleatórias com a correção da variável resposta por meio da subamostragem. Com base nisso, serão incorporadas as variáveis do Auxílio Emergencial a este modelo, com o objetivo de avaliar o impacto dessas informações, provenientes de dados públicos, no aprimoramento do modelo de concessão de crédito. A Tabela 5.10 a seguir, resume este comparativo.

Tabela 5.10: Comparativo dos resultados do modelo selecionado

Modelo	Base	Acurácia	Precisão	Sensibilidade	F1-Score	KS	Especificidade	G-mean	AUC-ROC	Kappa
Subamostragem										
Florestas Aleatórias	Treino	0.536	0.529	0.652	0.584	0.073	0.420	0.274	0.554	0.072
	Teste	0.467	0.247	0.646	0.357	0.052	0.419	0.266	0.536	0.031
	K-fold	0.536	0.527	0.648	0.582	0.061	0.421	0.274	0.554	0.073
Subamostragem + Variáveis do Auxílio Emergencial										
Florestas Aleatórias	Treino	0.561	0.576	0.464	0.514	0.123	0.658	0.305	0.587	0.122
	Teste	0.599	0.554	0.455	0.507	0.108	0.642	0.292	0.567	0.079
	K-fold	0.562	0.560	0.454	0.509	0.123	0.670	0.305	0.589	0.125

Ao comparar os resultados dos dois modelos — Florestas Aleatórias com subamostragem e Florestas Aleatórias com subamostragem mais a inclusão das variáveis do Auxílio Emergencial — é possível observar ganhos significativos em diversos aspectos, especialmente no desempenho do KS quando as novas variáveis foram incorporadas.

No conjunto de treino, a inclusão das variáveis do Auxílio Emergencial resultou em um aumento na Precisão, que passou de 0.529 para 0.576. Embora o F1-score tenha apresentado uma leve queda, a AUC subiu de 0.554 para 0.587, sugerindo que o modelo está se tornando mais robusto na capacidade de discriminação entre as classes, mesmo com essa ligeira queda no F1-score. Outro ponto de destaque foi a especificidade, que aumentou de 0.420 para 0.658, demonstrando que o modelo conseguiu captar mais casos verdadeiros de inadimplência.

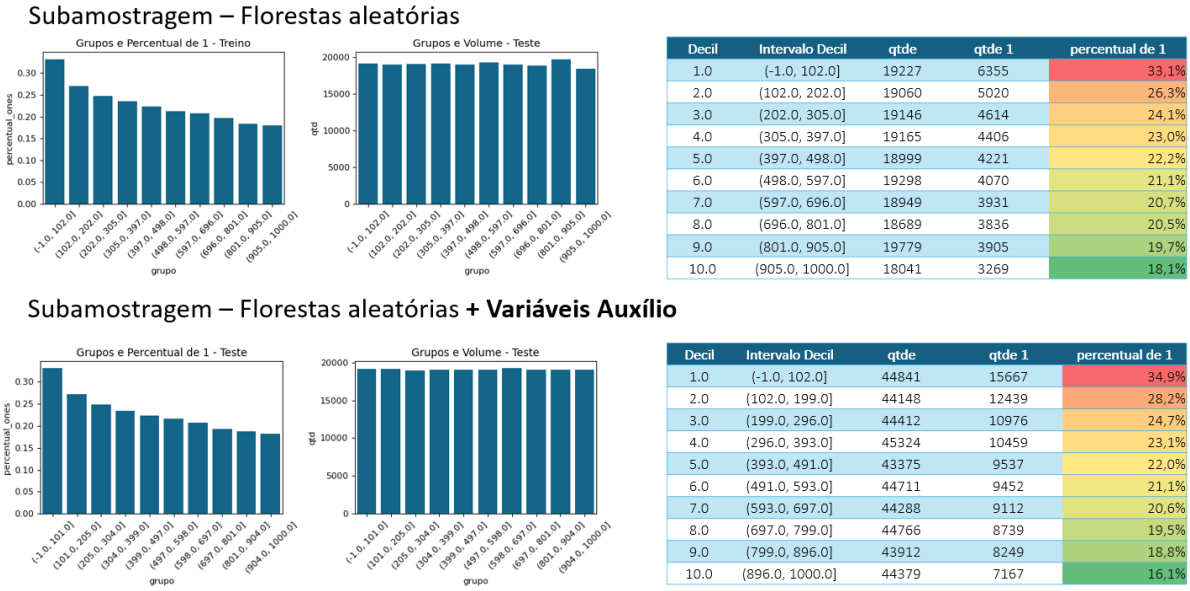
No conjunto de teste, os ganhos se tornaram ainda mais evidentes. O KS aumentou de 0.052 para 0.108, indicando uma melhor separação entre as classes. A Precisão passou de 0.247 para 0.554, mostrando que o modelo com as novas variáveis está sendo mais eficaz na classificação correta dos clientes. O F1-score também apresentou uma melhora, subindo de 0.357 para 0.507. A especificidade, por sua vez, subiu de 0.419 para 0.642, reforçando a capacidade do modelo em identificar clientes inadimplentes.

Destaca-se que, ao acrescentar as informações do Auxílio Emergencial, houve uma redução da sensibilidade, saindo de 0.646 para 0.455 na base de teste. Isso implica que houve uma redução na identificação dos bons pagadores, ou seja, uma quantidade menor de bons pagadores não será aprovada. Contudo, vale salientar que, neste trabalho é mais importante identificar os casos mais prováveis de inadimplência.

Em suma, a adição dessas variáveis trouxe ganhos expressivos, especialmente nos indicadores de Precisão, KS, e especificidade, que são fundamentais para melhorar a acurácia e eficiência na detecção de inadimplentes. Esses resultados sugerem que as informações do

Auxílio Emergencial foram úteis para refinar a segmentação e aumentar o poder preditivo do modelo de crédito.

O impacto positivo da inclusão das variáveis do Auxílio Emergencial pode ser observado na Figura 5.10, que compara a ordenação dos decis e os níveis de inadimplência antes e depois da incorporação dessas informações. Um dos principais destaques é o aumento da concentração de inadimplência no primeiro decil, que subiu de 33,1% para 34,9%, indicando uma melhora na capacidade do modelo em identificar os clientes de maior risco. Além disso, o melhor decil apresentou uma redução significativa na inadimplência, passando de 18,1% para 16,1%, evidenciando o ganho de performance na classificação dos melhores clientes após a adição dessas variáveis.



Decil	Intervalo Decil	qtde	qtde 1	percentual de 1
1.0	(-1.0, 102.0]	19227	6355	33,1%
2.0	(102.0, 202.0]	19060	5020	26,3%
3.0	(202.0, 305.0]	19146	4614	24,1%
4.0	(305.0, 397.0]	19165	4406	23,0%
5.0	(397.0, 498.0]	18999	4221	22,2%
6.0	(498.0, 597.0]	19298	4070	21,1%
7.0	(597.0, 696.0]	18949	3931	20,7%
8.0	(696.0, 801.0]	18689	3836	20,5%
9.0	(801.0, 905.0]	19779	3905	19,7%
10.0	(905.0, 1000.0]	18041	3269	18,1%

Decil	Intervalo Decil	qtde	qtde 1	percentual de 1
1.0	(-1.0, 102.0]	44841	15667	34,9%
2.0	(102.0, 199.0]	44148	12439	28,2%
3.0	(199.0, 296.0]	44412	10976	24,7%
4.0	(296.0, 393.0]	45324	10459	23,1%
5.0	(393.0, 491.0]	43375	9537	22,0%
6.0	(491.0, 593.0]	44711	9452	21,1%
7.0	(593.0, 697.0]	44288	9112	20,6%
8.0	(697.0, 799.0]	44766	8739	19,5%
9.0	(799.0, 896.0]	43912	8249	18,8%
10.0	(896.0, 1000.0]	44379	7167	16,1%

Figura 5.10: Ordenação em decis do modelo final.

Foi realizado ainda o gráfico de importância das variáveis e o gráfico SHAP (*SHapley Additive Explanations*), ambos fornecem informações do impacto das variáveis no desempenho do modelo.

No gráfico de importância das variáveis (Figura 5.11), é possível observar quais atributos mais contribuíram para a tomada de decisão do modelo. Neste caso, as variáveis mais importantes foram as demográficas. Em primeiro lugar ficou a variável idade do cliente seguido da variável que descreve se o cliente é da região nordeste ou não e se ele está solteiro ou não. A primeira informação do Auxílio Emergencial ficou em 11º de um total de 30 variáveis no modelo final.

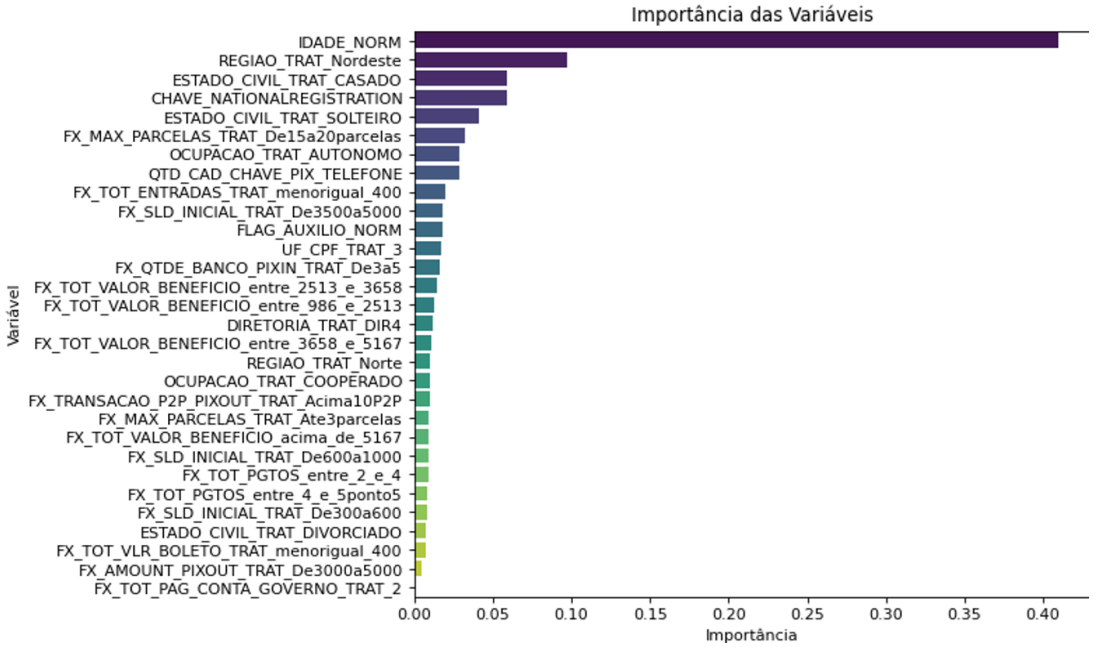


Figura 5.11: Importância das variáveis.

Já os valores da Figura 5.12 de SHAP complementam essa análise ao mostrar o impacto individual de cada variável nas previsões. Com os valores de SHAP, podemos entender não apenas quais variáveis são importantes, mas também se o efeito delas é positivo ou negativo para a classificação de risco de crédito. A análise visual dos valores de SHAP revela como as variáveis do Auxílio Emergencial, como o recebimento ou não do benefício, influenciam diretamente a probabilidade de inadimplência, oferecendo uma visão interpretável de como o modelo utiliza esses dados para fazer suas previsões. Isso é crucial para garantir que o modelo não apenas seja mais preciso, mas também mais transparente e confiável.

O eixo horizontal representa o valor SHAP, que indica o impacto de cada variável na previsão do modelo. Valores positivos indicam que a variável aumenta a probabilidade do resultado positivo (por exemplo, a aprovação de crédito), enquanto valores negativos sugerem que a variável diminui essa probabilidade.

As cores dos pontos variam do azul ao vermelho. A cor azul indica valores baixos para a variável, enquanto a cor vermelha representa valores altos. Isso ajuda a visualizar rapidamente se uma variável com um valor específico está contribuindo positivamente ou negativamente para a previsão. Por fim, a densidade e a distribuição dos pontos ao longo do eixo horizontal mostram como cada variável se comporta em relação à previsão. Uma distribuição ampla sugere que a variável tem um impacto significativo na previsão,

enquanto uma distribuição mais concentrada pode indicar que a variável tem um efeito mais limitado.

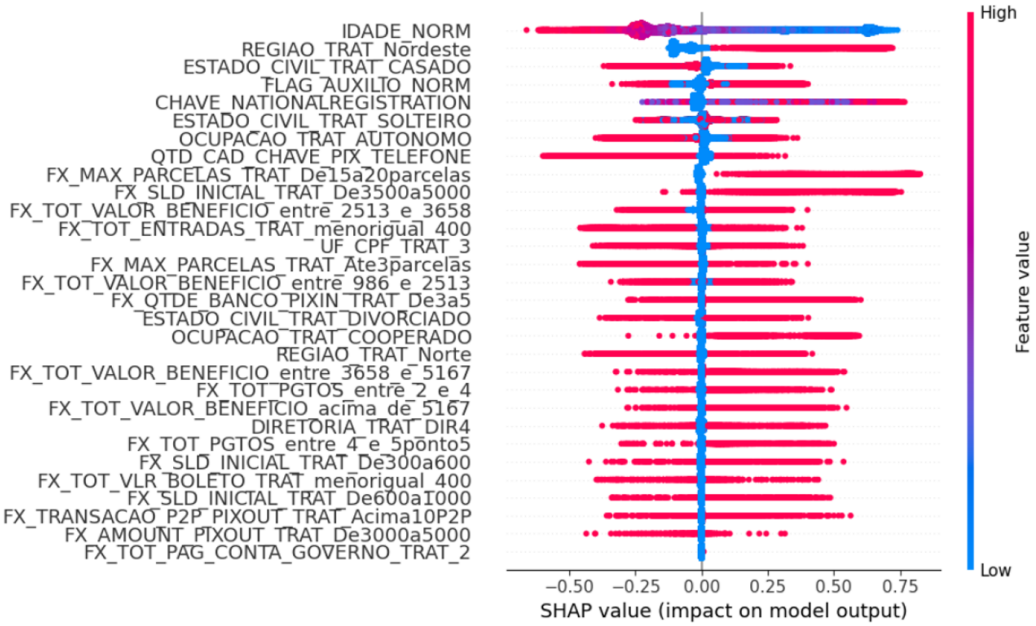


Figura 5.12: Gráfico Shap Values.

No caso da Figura 5.12, pode-se interpretar que a variável idade, por exemplo, tem um valor SHAP positivo significativo quando em um determinado intervalo (representado pela cor vermelha), isso sugere que, à medida que a idade aumenta, a probabilidade de aprovação de crédito tende a aumentar. No caso da variável REGIAO_TRAT_Nordeste os pontos estão mais concentrados à direita do gráfico, o que significa que, quem for da região nordeste pode ser favorecido no momento da oferta do crédito.

Uma análise final do modelo foi realizada, incorporando a informação de que os clientes pertenciam a um dos seguintes estados: São Paulo, Espírito Santo, Rio de Janeiro, Rio Grande do Sul e Paraná. Essas informações foram sintetizadas na variável SERRP, que assumia o valor 1 caso o cliente fosse residente em um desses estados. Ao rodar o modelo com a inclusão dessa variável, observou-se que ela apresentou uma importância significativa, alcançando a nona posição no gráfico de importância das variáveis, conforme mostra a Figura 5.13. Esse resultado sugere que a localização geográfica dos clientes exerce influência relevante sobre a decisão de concessão de crédito.

A hipótese levantada é de que clientes residentes nesses estados podem ter um perfil econômico mais favorável ou acesso a melhores condições de mercado, o que aumenta sua probabilidade de aprovação de crédito. Essa descoberta abre espaço para análises

regionais mais aprofundadas, uma vez que as diferenças econômicas e sociais entre os estados podem afetar diretamente o comportamento de crédito dos indivíduos.

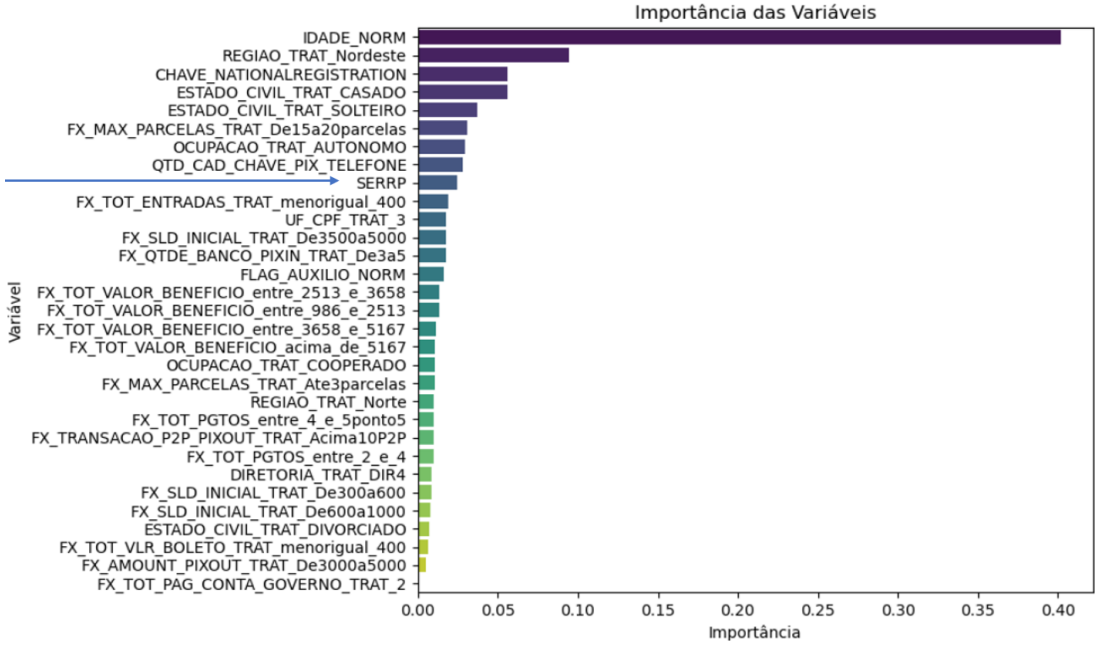


Figura 5.13: Gráfico de importância com SEERP.

Após a definição do modelo e a análise de performance, o time de modelagem encaminha o modelo ao time de políticas de crédito para uma avaliação adicional. Esse time tem a responsabilidade de criar as regras práticas de concessão de crédito, com base no modelo desenvolvido e em outras informações estratégicas e regulatórias. A partir da análise do modelo e levando em conta dados adicionais, como perfis de risco e limites operacionais, o time de políticas define, por exemplo, a segmentação dos clientes em diferentes grupos de risco, categorizando-os desde os menos arriscados até os mais propensos à inadimplência. Além disso, o time de políticas pode ajustar limites de crédito e definir ações específicas para cada faixa de risco, visando maximizar a eficácia das decisões de crédito e a sustentabilidade da carteira, sempre alinhado aos objetivos estratégicos da instituição.

Capítulo 6

Conclusão

O presente estudo analisou a eficácia de diferentes modelos de aprendizado de máquina na previsão da inadimplência de clientes de uma instituição financeira, com foco particular na influência das variáveis relacionadas ao Auxílio Emergencial sobre o desempenho desses modelos de crédito. A partir da aplicação de diversas técnicas de aprendizado de máquina e de métodos para correção do desbalanceamento da variável resposta, observou-se que a técnica de Florestas Aleatórias, combinada com subamostragem, apresentou ganhos significativos em acurácia e na capacidade de classificação ao incorporar informações referentes ao Auxílio Emergencial.

A análise dos resultados mostrou que os modelos de Florestas Aleatórias foram os mais eficazes quando combinados com técnicas de balanceamento, como SMOTE e subamostragem. Isso se deve à natureza dessas técnicas, que permitem capturar melhor as interações entre variáveis e ajustar o aprendizado para a classe minoritária.

O modelo de Florestas Aleatórias apresentou uma ordenação aprimorada dos decis, evidenciada pelo aumento do percentual de inadimplência do primeiro decil, que passou de 33,1% para 34,9%. O último decil também demonstrou uma redução na inadimplência, caindo de 18,1% para 16,1%. Essa mudança reflete uma maior capacidade do modelo em identificar padrões de risco com a inclusão das variáveis do Auxílio. A análise dos valores SHAP revelou que a variável FLAG_AUXILIO_NORM teve um impacto considerável nas previsões do modelo. Valores altos dessa variável mostraram uma tendência de aumentar a probabilidade de inadimplência, destacando a importância de considerar a situação financeira dos beneficiários de auxílio na avaliação de crédito.

A comparação dos resultados com e sem as variáveis do Auxílio Emergencial demonstrou que a inclusão dessas informações não apenas melhorou a precisão do modelo, mas

também proporcionou uma compreensão mais aprofundada dos fatores que influenciam a concessão de crédito. O ganho em métricas como precisão, especificidade, KS e F1-score, especialmente no cenário com Florestas Aleatórias, indicou uma maior robustez na avaliação do risco de crédito.

A inclusão das informações do Auxílio Emergencial no processo de concessão de crédito permite uma abordagem mais informada e responsiva às condições econômicas dos indivíduos, especialmente em se tratando de uma parcela da população que normalmente não consegue acesso a crédito nos bancos tradicionais. Isso não apenas melhora o desempenho dos modelos, mas também promove práticas de concessão de crédito mais justas e alinhadas com as realidades financeiras dos clientes. Além disso, a combinação das variáveis internas com as informações do Auxílio Emergencial permite elaborar um modelo que considera as características específicas das pessoas das classes C, D e E, frequentemente desconsideradas nos modelos de crédito tradicionais.

Uma das limitações deste estudo é que ele se baseou em uma amostra representativa da carteira de clientes da instituição, e não a base completa. Além disso, as informações utilizadas no modelo foram aquelas disponíveis no portal de dados à época da análise considerando o período de pandemia, limitando a incorporação de outras variáveis que poderiam aumentar a capacidade preditiva do modelo. Essas restrições, no entanto, são inerentes a um processo contínuo de desenvolvimento e aprimoramento dos modelos de crédito, e refletem o estágio de maturidade dos dados disponíveis no momento da pesquisa.

Recomenda-se que futuros estudos aprofundem a análise da relação entre variáveis socioeconômicas e o risco de crédito, utilizando técnicas de modelagem mais sofisticadas e explorando novas fontes de dados que possam enriquecer a capacidade preditiva dos modelos. Além disso, é fundamental realizar um acompanhamento contínuo da performance do modelo após a inclusão das variáveis relacionadas ao Auxílio Emergencial, garantindo que a avaliação de risco se mantenha precisa e alinhada às mudanças nas condições econômicas e sociais dos clientes. Dessa forma, a integração dessas variáveis não só demonstrou uma melhoria significativa no desempenho preditivo, como também destacou a importância de uma abordagem mais personalizada e contextualizada na análise de crédito, adaptando-se às especificidades desse grupo populacional.

Em conclusão, este estudo ressalta a importância de considerar variáveis externas, como o Auxílio Emergencial e demais benefícios sociais, para aprimorar a análise de crédito em populações vulneráveis. A incorporação dessas informações torna os modelos mais eficazes e personalizados, refletindo melhor as realidades econômicas dos clientes. Futuras

pesquisas e ajustes constantes serão essenciais para garantir que os modelos de crédito continuem evoluindo, garantindo práticas mais inclusivas e adequadas ao cenário econômico em transformação.

Referências

[1] Forti, Melissa: *Técnicas de machine learning aplicadas na recuperação de crédito do mercado brasileiro*, 2018. Dissertação (MPFE) - Escola de Economia de São Paulo. ix, 3, 16, 23, 24, 27

[2] Gonzalez, Leandro de Azevedo: *Regressão logística e suas aplicações*. Monografia (Graduação) – Universidade Federal do Maranhão, Centro de Ciências Exatas e Tecnológicas, Curso de Graduação em Ciência da Computação., 2018. ix, 24, 26

[3] Sicsú, Abraham Laredo: *Credit Scoring: desenvolvimento, implantação, acompanhamento*, volume 1. São Paulo, Blucher, 2010. ix, 16, 22, 23, 34, 36, 37, 38

[4] Azevedo, Ana e Manuel Filipe Santos: *KDD, SEMMA and CRISP-DM: A Parallel Overview*. Single, ISBN 978-972-8924-63-8, 2008. ix, 39, 40

[5] *Síntese de Indicadores Sociais - Uma Análise das Condições de Vida da População Brasileira*. Instituto Brasileiro de Geografia e Estatística (IBGE), 2021. `https://biblioteca.ibge.gov.br/visualizacao/livros/liv101892.pdf`. 1

[6] Brasileiro, Governo: *Portal de dados abertos brasileiro*. `https://www.gov.br/governodigital/pt-br/dados-abertos/portal-brasileiro-de-dados-abertos`, acesso em 2022-05-05. 1, 24

[7] Brasil, Banco Central do: *Sistema financeiro e instituições financeiras*. `https://www.bcb.gov.br/estabilidadefinanceira/bancoscaixaseconomicas`, acesso em 2022-06-05. 2

[8] Brasil, Banco Central do: *Fintechs de crédito e bancos digitais*. Estudo Especial nº 89/2020 – Divulgado originalmente como boxe do Relatório de Economia Bancária (2019), 89, 2020. 2

[9] Moscato, Vicenzo, Antonio Picariello e Giancarlo Sperlí: *A benchmark of machine learning approaches for credit score prediction*. Expert Systems with Applications, 165:113986, 2021. 2, 33, 34, 35, 36, 37

[10] Pisani, Paulo Hnerique e Ana Carolina Lorena: *A systematic review on keystroke dynamics*. Journal of the Brazilian Computer Society, 9(4):573–597, 2013. 6

[11] Kitchenham, Barbara, O. Pearl Brereton, David Budgen, Mark Turner, John Bailey e Stephen Linkman: *Systematic literature reviews in software engineering – a systematic literature review*. Information and Software Technology, 51:7–15, 2009. 6

[12] Milian, Eduardo Z, Mauro de M Spinola e Marly M de Carvalho: *Fintechs: A literature review and research agenda.* Electronic Commerce Research and Applications - Elsevier, 34:100833, 2019. 7

[13] Bussmann, Niklas, Paolo Giudici, Dimitri Marinelli e Jochen Papenbrock: *Explainable ai in fintech risk management.* Frontiers in Artificial Intelligence, 3:26, 2020. 16

[14] Leong, Carmen, Barneyb Tan, Xiao Xiao, Felix Ter Chian Tan e Yuan Sun: *Nurturing a fintech ecosystem: The case of a youth microloan startup in china.* International Journal of Information Management, 37:92–97, 2017. 17

[15] Giudici, Paolo, Branka Hadji-Misheva e Alessandro Spelta: *Network based scoring models to improve credit risk management in peer to peer lending platforms.* Frontiers in Artificial Intelligence, 2:3, 2019. 17

[16] Najib, Mukhamad, Wita Juwita Ermawati, Farah Fahma, Endri Endri e Dwi Suhartanto: *Fintech in the small food business and its relation with open innovation.* Journal of Open Innovation: Technology, Market, and Complexity, 7:88, 2021. 17

[17] Appiah-Otoo, Isaac e Na Song: *The impact of fintech on poverty reduction: Evidence from china.* Sustainability, 13:5225, 2021. 17

[18] Kim, Cho: *Towards repayment prediction in peer-to-peer social lending using deep learning.* Mathematics, 7:1041, 2022. 17

[19] Yan, Jiaqi, Wayne Yu e J. Leon Zhao: *How signaling and search costs affect information asymmetry in p2p lending: the economics of big data.* Financial Innovation, 1:19, 2015. 17, 30

[20] Ahelegbey, D. F., Paolo Giudici e Branka Hadji-Misheva: *Latent factor models for credit scoring in p2p systems.* Physica A, 2019. 17

[21] Bernards, Nick: *The poverty of fintech? psychometrics, credit infrastructures, and the limits of financialization.* Review of International Political Economy, 2022. 18

[22] Demirgüç-Kunt, Asli, Leora Klapper, Dorothe Singer, Saniya Ansar e Jake Hess: *The global findex database 2017: Measuring financial inclusion and opportunities to expand access to and use of financial services.* The World Bank Economic Review, 34:S2–S8, 2022. 18

[23] Alcantara, Williams, Judson Bandeira, Armando Barbosa, André Lima, Thiago Ávila, Igor Bittercourt e Seiji Isotani: *Desafios no uso de dados abertos conectados na educação brasileira.* SBC, 1:11–20, 2015. 18

[24] Pavão, Caterina Groposo, Rafael Porte da Rocha e Rene Faustino Gabriel Junior: *Proposta de criação de uma rede de dados abertos da pesquisa brasileira.* RDBCI: Rev. Digit. Bibliotecon. Cienc. Inf., 16:329–343, 2018. 18

[25] Santos Neto, A.L., C.H. Marcondes, D.V. Pereira, E.R. Fonseca, I.V.P. Souza, Nilson Barborsa, R.P.T. Moraes e S.C. Martins: *Tecnologias de dados abertos para interligar bibliotecas, arquivos e museus: um caso machadiano.* TransInformação, Campinas, 25:81–87, 2013. 18

[26] Neves, José Anael, Mick Lennon Machado, Luna Dias de Almeida Oliveira, Yara Maria Franco Moreno, Maria Angélica Tavares de Medeiros e Francisco de Assis Guedes de Vasconcelos: *Unemployment, poverty, and hunger in brazil in covid-19 pandemic times*. Rev Nutr, 34, 2021. 19

[27] Feng, Xiaodong, Zhi Xiao, Bo Zhong, Jing Qiu e Yuanxiang Dong: *Dynamic ensemble classification for credit scoring using soft probability*. Elsevier - Applied Soft Computing, 65:139–151, 2018. 19

[28] Soui, Makram, Ines Gasmi, Salima Smiti e Khaled Ghédira: *Rule-based credit risk assessment model using multi-objective evolutionary algorithms*. Expert Systems with Applications, 126:144–157, 2019. 19

[29] García, Vicente, Ana I. Marqués e J. Salvador Sánchez: *Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction*. Elsevier, 47:88–101, 2019. 19

[30] Abellán, Joaquín e Javier G. Castellano: *A comparative study on base classifiers in ensemble methods for credit scoring*. Expert Systems with Applications, 73:1–10, 2017. 19

[31] Abellán, Joaquín e Carlos J. Mantas: *Improving experimental studies about ensembles of classifiers for bankruptcy prediction and credit scoring*. Expert Systems with Applications, 41:3825–3830, 2014. 20

[32] Xia, Yufei, Junhao Zhao, Lingyun He, Yinguo Li e Mengyi Niu: *A novel tree-based dynamic heterogeneous ensemble method for credit scoring*. Expert Systems with Applications, 159:113615, 2020. 20, 29, 30

[33] Bequé, Artem e Stefan Lessmann: *Extreme learning machines for credit scoring: An empirical evaluation*. Expert Systems with Applications, 86:42–53, 2017. 20

[34] Wang, Lu, Yuangao Chen, Hui Jiang e Jianrong Yao: *Imbalanced credit risk evaluation based on multiple sampling, multiple kernel fuzzy self-organizing map and local accuracy ensemble*. Applied Soft Computing, 91:106262, 2020. 20, 34, 35, 36

[35] Sun, Jie, Jie Lang, Hamido Fujita e Hui Li: *Imbalanced enterprise credit evaluation with dte-sbd: Decision tree ensemble based on smote and bagging with differentiated sampling rates*. Information Sciences, 425:76–91, 2018. 20, 33, 34, 35, 36

[36] Abe, Naoki, Bianca Zadronzy e John Langford: *An iterative method for multi-class cost-sensitive learning*. Association for Computing Machinery, páginas 3–11, 2004. 21

[37] Yin, Qing Yan, Jiang She Zhang, Chun Xia Zhang e Sheng Cai Liu: *An empirical study on the performance of cost-sensitive boosting algorithms with different levels of class imbalance*. Hindawi Publishing Corporation Mathematical Problems in Engineering, 2013:12, 2013. 21

[38] Krawczyk, Bartosz: *Cost-sensitive one-vs-one ensemble for multi-class imbalanced data*. 2016 International Joint Conference on Neural Networks (IJCNN), páginas 2447–2452, 2016. 21, 30

[39] Brasil, Banco Central do: *Resolução nº 4.557, de 23 de fevereiro de 2017*. `https://normativos.bcb.gov.br/Lists/Normativos/Attachments/50344/Res_4557_v1_O.pdf`, acesso em 2023-03-20. 22

[40] Neto, Jonas Arruda Novaes: *Modelo preditivo de capacidade de pagamento para prospecção pf: Atraindo e fidelizando clientes no cenário de open finance*, 2022. Dissertação (Mestrado) - Universidade de Brasília - UNB. 23

[41] David W. Hosmer, Jr., Stanley Lemeshow e Rodney X. Sturdivant: *Applied Logistic Regression*, volume 3rd ed. John Wiley and Sons, Hoboken, New Jersey, 2013. 24

[42] Morettin, Pedro Alberto e Julio Motta: *Estatística e Ciência de Dados*, volume 1st ed. Singer ISBN 978-85-216-3816-2, 2022. 28, 38, 46, 47, 56

[43] Breiman, Leo: *Random forests*. Kluwer Academic Publishers, 45:5–32, 2001. 28

[44] Ampomah, Ernest Kwame, Zhiguang Qin e Gabriel Nyame: *Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement*. MDPI, Journal Information, 11:332, 2020. 28

[45] Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye e Tie Yan Liu: *Lightgbm: A highly efficient gradient boosting decision tree*. Advances in Neural Information Processing Systems, Volume 2017-December:3147–3155, 2017. 29, 30

[46] Taha, Altyeb Altaher e Sharaf Jameel Malebary: *An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine*. IEEE Access, 8:25579–25587, 2020. 30

[47] Xiong, Yueling, Mingquan Ye e Changrong Wu: *Cancer classification with a cost-sensitive naive bayes stacking ensemble*. Computational and Mathematical Methods in Medicine, 2021:12, 2021. 30, 31

[48] Di Nunzio, Giorgio Maria: *A new decision to take for cost-sensitive naïve bayes classifiers*. Information Processing and Management, 50:653–674, 2014. 30

[49] Shan, Wangweiyi, Chao Dong, Yi Wu e Xuhua Pan: *A random feature mapping method based on the adaboost algorithm and results fusion for enhancing classification performance*. Information Sciences, 644:315–331, 2023. 31, 33

[50] Junior, Guanis B. Vilela, Bráulio N. Lima, Adriano A. Pereira, Marcelo F. Rodrigues, José Ricardo L. Oliveira, Luís F. Silio, Anderson S. Carvalho, Heros R. Ferreira e Ricardo P. Passos: *Métricas utilizadas para avaliar a eficiência de classificadores em algoritmos inteligentes*. Centro de Pesquisas Avançadas em Qualidade de Vida, 14, 2022. 37

[51] Python: *Optuna: Uma estrutura de otimização de hiperparâmetros.* `https://optuna.readthedocs.io/en/stable/`, acesso em 2023-06-05. 38

[52] Inc., Databricks: *Databricks platform.* `https://www.databricks.com/`, acesso em 2023-06-05. 41

[53] Copyright 2007 2024, scikit learn developers (BSD License).: *Sequentialfeatureselector.* `https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SequentialFeatureSelector.html`, acesso em 2023-06-05. 47

[54] Andrew, Y. Ng: *Feature selection, l1 vs. l2 regularization, and rotational invariance.* Appearing in Proceedings of the 21 st International Conference on Machine Learning, Banff, Canada, páginas 1–8, 2004. 50

[55] Copyright 2007 2024, scikit learn developers (BSD License).: *Logisticregression.* 50

[56] Il-Gyo, Chong e Jun Chi-Hyuck: *Performance of some variable selection methods when multicollinearity is present.* Elsevier - Chemometrics and Intelligent Laboratory Systems, 78:103–112, 2005. 56

[57] Kuhn, M. e K. Johnson: *Applied predictive modeling. Springer.* Springer, ISBN: 978-1461468486, 2013. 56